Domain Adaptive Transfer Learning for Visual Classification

by

ASHIQ IMRAN

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2021

Domain Adaptive Transfer Learning for Visual Classification

The members of the Committee approve the doctoral
dissertation of Ashiq Imran

Vassilis Athitsos
Supervising Professor ____________________

Farhad Kamangar ____________________

Christopher Conly ____________________

David Levine ____________________

Dean of the Graduate School ____________________

Copyright © by Ashiq Imran 2021

All Rights Reserved

ACKNOWLEDGEMENTS

First and foremost, I am thankful to the almighty for giving me the patience, perseverance and strength required to accomplish this goal of gaining and completing my Ph.D. degree. I like to thank my parents, my wife, my brother, my sister-in-law, my nephews, my friends and all my well wishers.

I am genuinely thankful to my supervising professor Dr. Vassillis Athitsos, who I am honored to be one of his Ph.D. students. I highly respect and appreciate his consistent help, support, guidance, patience, understanding, encouragement, and insightful advice during my entire time in my doctoral journey. I would also like to thank my academic committee members: Dr. Farhad Kamangar, Professor David Levine, and Dr. Christopher Conly, for their interest, assistance, and valuable feedback.

July 29, 2021

ABSTRACT

Domain Adaptive Transfer Learning for Visual Classification

Ashiq Imran, Ph.D. Candidate

The University of Texas at Arlington, 2021

Supervising Professor: Vassilis Athitsos

Deep Neural Networks have made a significant impact on many computer vision applications with large-scale labeled datasets. However, in many applications, it is expensive and time-consuming to gather large-scale labeled data. With the limited availability of labeled data, it is challenging to obtain great performance. Moreover, in many real-world problems, transfer learning has been applied to cope with limited labeled training data. Transfer learning is a machine learning paradigm where pre-trained models on one task can be reused for another task. This dissertation investigates transfer learning and related machine learning techniques such as domain adaptation on visual categorization applications.

At first, we leverage transfer learning on fine-grained visual categorization (FGVC). FGVC is a challenging topic in computer vision. FGVC is different from general recognition. It is a problem characterized by large intra-class differences and subtle inter-class differences. FGVC should be capable of recognizing and localizing the nuances within subordinate categories. We tackle this problem in a weakly supervised manner, where neural network models are getting fed with additional data using a data augmentation technique through a visual attention mechanism. We perform domain adaptive knowledge transfer via fine-tuning on our base network model. We perform our experiment on six challenging and commonly used FGVC datasets. We show competitive improvement on accuracy by using attention-aware data augmentation techniques with features derived from the deep learning model InceptionV3, pre-trained on large-scale datasets. Our method outperforms competitor methods on multiple FGVC datasets and showed competitive results on other datasets. Experimental studies show that transfer learning from large-scale datasets can be utilized effectively with visual attention-based data augmentation, obtaining state-of-the-art results on several FGVC datasets.

In many applications, specifically for transfer learning, it is assumed that the source and target domain have the same distribution. However, it is hardly true in real-world applications. Moreover, direct transfer across domains often performs poorly because of domain shift. Domain adaptation, a sub-field of transfer learning, has become a prominent problem setting that refers to learning a model from a source domain that can perform reasonably well on the target domain. This dissertation investigates and proposes improvements on visual categorizations using domain adaptation. Following the context of domain adaptation, a literature review covering and summarizing the most recently proposed domain adaptation method is presented. Finally, we propose a technique that uses the adaptive feature norm with subdomain adaptation to boost the transfer gains. Subdomain adaptation can enhance the ability of deep adaptation networks by capturing the fine-grained features from each category. Additionally, we have incorporated an adaptive feature norm approach to increase transfer gains. Our method shows state-of-the-art results on standard cross-domain adaptation datasets for the object categorization task.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Deep Neural Networks have shown remarkable performance in various domains in the field of computer vision. In order to achieve good performance, it requires a humongous amount of labeled data. Training larger and deeper networks are difficult, particularly if the size of a dataset is small. Additionally, collecting well-annotated data is costly and time-consuming. In computer vision, transfer learning, popularly known as cross-domain adaptation [6], aims at enabling the generalization capacity of the target domain by loading knowledge from the source domain. Typically, a way to regularize deep networks is to simply use a pre-trained model which has been trained on a different dataset and use this model for the target dataset [7]. However, fine-tuning still requires a considerable amount of labeled training data, which may not be available for many applications.

Visual recognition is one of the most fundamental problems in the field of computer vision. A computer vision system requires the generalization of many object variations due to viewpoint, illumination, or occlusions. Additionally, it needs to be specific towards recognizing objects.

One of the most challenging problems in the field of object recognition is Fine-Grained Visual Categorization (FGVC). In FGVC, same-class items may have variations in the pose, scale, or rotation. FGVC contains subtle differences among classes in a sub-category of an object, which includes the model of the cars, type of the foods or the flowers, species of the birds or dogs, and type of the aircraft. These differences make FGVC a challenging problem, as there are significant intra-class differences among the sub-categories, and at the same time, items from different classes may look similar. In contrast with traditional object classification techniques, FGVC aims to solve the identification of particular sub-categories from a given category [8, 1]. A domain adaptive knowledge transfer to select the optimal source domain for the target FGVC domains is performed. Then, we apply a visual attention-aware data augmentation technique to boost up the FGVC accuracy. This research leads us to investigate further in the field of domain adaptive transfer learning.

Typically, in transfer learning [9], a pre-trained model on a large dataset is used to fine-tune the target dataset. This method may not generalize well to new environments and new datasets. One of the reasons is that deep learning methods assume that training and testing data are drawn from independent and identical distributions (i.i.d). However, this assumption rarely holds, as there will be a shift in data distributions across different domains. Furthermore, traditional machine-learning paradigms like supervised learning train models to predict the outcome for unseen data. These models do not necessarily optimize performance if the difference between the test and training data is great. Domain adaptation can be a way to mitigate these issues and reduce the effort of recollecting and retraining a model by transferring knowledge between tasks and domains [10]. Domain adaptation is a technique that alleviates networks' performance degradation on the target samples, which has a different distribution from the source samples [3]. Since most existing domain adaptation methods assume that source and target domain contain the same labels, performance degradation occurs when the label set of the source and target are not the same. It is challenging to use existing domain adaptation techniques in real situations because generating the source domain with exactly the same label set as the target domain is ineffective. Unsupervised domain adaptation methods have been attended to solve this problem, which intends to learn this latent space by aligning data across domains.

This dissertation will go over in detail of transfer learning and domain adaptation. The dissertation started by analyzing the current state of the problem. Then, it proposes a new technique of domain adaptive transfer learning on fine-grained visual categorization. A detailed survey on unsupervised domain adaptation is presented afterward. Subsequently, a novel approach of considering sub-domains with adaptive feature norms is described. The efficacy of the proposed method has been tested on several datasets and compared with state-of-the-art methods.

1.1 Dissertation Contributions

The focus of this dissertation is on domain adaptation in visual classification tasks. The work presented in the following chapters will make the following contributions:

1. Domain Adaptive Transfer Learning on Visual Attention Aware Data Augmentation for Fine-grained Visual [11] Categorization.
2. A comprehensive literature survey on unsupervised domain adaptation on visual classification.
3. Adaptive Feature Norm for Unsupervised Sub-Domain Adaptation.

1.2 Dissertation Organization

Chapter 2 aims to tackle fine-grained visual classification with a domain adaptive transfer learning approach. Our proposed method achieves state-of-the-art results in multiple fine-grained classification datasets, including CUB200_2011 bird datasets, Flowers-102, and FGVC-Aircrafts.

In chapter 3, it covers a comprehensive literature survey of unsupervised domain adaptation, including discrepancy-based methods, adversarial methods, and reconstruction-based methods. A comparison of the neural network-based domain adaptation methods is summarized.

In chapter 4, A technique is proposed that uses the adaptive feature norm with subdomain adaptation to boost up the transfer gains. Subdomain adaptation can enhance the ability of deep adaptation networks by capturing the fine-grained features from each category. Additionally, we have incorporated an adaptive feature norm approach to increase transfer gains. Our method shows state-of-the-art results on the popular visual classification datasets, including Office-31, Office Home, and Image-CLEF datasets.

Chapter 5 discusses some of the research directions of the unsupervised domain adaptation.

In Chapter 6, the summary of contributions of my research is presented.

CHAPTER 2

Domain Adaptive Transfer Learning on Visual Attention Aware Data Augmentation for Fine-grained Visual Categorization

2.1 Introduction

Deep neural networks have provided state-of-the-art results in many domains in computer vision. However, having a big training set is very important for the performance of deep neural networks [12, 13]. Data augmentation techniques have been gaining popularity in deep learning and are extensively used to address the scarcity of training data. Data augmentation has led to promising results in various computer vision tasks [13]. There are different data augmentation methods for deep models, like image flipping, cropping, scaling, rotation, translation, color distortion, adding Gaussian noise, and many more.

Previous methods mostly choose random images from the dataset and apply the above operations to enlarge the amount of training data. However, applying random cropping to generate new training examples can have undesirable consequences. For example, if the size of the cropped region is not large enough, it may consist entirely of background, and not contain any part of the labeled object. Moreover, this generated data might reduce accuracy and negatively affect the quality of the extracted features. Consequently, the disadvantages of random cropping might cancel out its advantages. More specific features need to be provided to the model to make data augmentation more productive.

In Fine-Grained Visual Categorization (FGVC), same-class items may have variation in the pose, scale, or rotation. FGVC contains subtle differences among classes in a sub-category of an object, which includes the model of the cars, type of the foods or the flowers, species of the birds or dogs, and type of the aircrafts. These differences are what make FGVC a challenging problem, as there are significant intra-class differences among the sub-categories, and at the same time, items from different classes may look similar. In contrast with regular object classification techniques, FGVC aims to solve the identification of particular subcategories from a given category [8, 1].

Convolutional Neural Networks (CNNs) have been extensively used for various applications in computer vision. To achieve good performance with CNNs, typically we need large amounts of labeled data. However, it is a tedious process to collect labeled fine-grained datasets. That is why there are not many FGVC datasets, and existing datasets are not as large compared to standard image recognition datasets like ImageNet [12]. Normally, a model pre-trained on large scale datasets such as ImageNet is used, and that model is then fine-tuned using data from an FGVC dataset. Typically, FGVC datasets are not too big, so it becomes critical to design methods that can compensate for the limited amount of data. In this paper, we investigate some techniques that allow the model to learn features more effectively, and that perform well on large scale datasets with fine-grained categories.

Generally, there are two domains involved in fine-tuning a network. One is the source domain, which typically includes large scale image datasets like ImageNet [12], where initial models are pre-trained. Another is the target domain, where data is used to fine-tune the pre-trained models. In this paper, the target domain is FGVC datasets, and we are interested in developing techniques that can boost accuracy on these type of datasets. Modern FGVC methods use pre-trained networks with ImageNet dataset to a large extent. We explore the possibility of achieving better accuracy than what has been achieved so far using ImageNet. A model first learns useful features from a large amount of training data, and is then fine-tuned on a more evenly-distributed subset to balance the efforts of the network among different categories and transfer the already learned features.

In short, our research tries to address two questions: 1) What approaches beyond transfer learning do we need to take to boost the performance on FGVC datasets? 2) How can we determine which large scale source domain we choose, given that the target domain is FGVC?

We calculate the domain similarity score between the source and target domains. This score gives us a clear picture of selecting the source domain for transfer learning to achieve better accuracy in the target domain. Then, we focus on a visual attention guided network for data augmentation. As FGVC datasets are relatively smaller in size, we leverage the feature learning from fine-tuning as well as data augmentation to achieve better accuracy. The performance of the combination of these two strategies outperforms the baseline approach.

In summary, the main contributions of this work are:

1. We propose a simple yet effective improvement over the recently proposed Weakly Supervised Data Augmentation Network (WS-DAN) [1], which is used for generating attention maps to extract sequential local features to tackle the FGVC challenge. A domain similarity score can play a vital role before applying transfer learning. Based on the score, we decide which source domain is necessary to use for transfer learning. Then, we can employ WS-DAN [1] to achieve better results among FGVC datasets.
2. We demonstrate a domain adaptive transfer learning approach, that combines with visual attention based data augmentation, and that can achieve state-of-the-art results on CUB200-2011 [14], and Flowers-102 [15], and FGVC-Aircrafts

[16] datasets. Additionally, we match the current state-of-the-art accuracy on Stanford Cars [17], Stanford Dogs [18] datasets.

3. We present the relationship of top-1 accuracy and domain score on six commonly used FGVC datasets. We illustrate the effect of image resolution in transfer learning in detail.

2.2 Related Work

In this section, we present a brief overview of data augmentation, fine-grained visual categorization, visual attention mechanism and transfer learning.

2.2.1 Data Augmentation

Machine learning theory suggests that a model can be more generalized and robust if it has been trained on a dataset with higher diversity. However, it is a very difficult and time-consuming task to collect and label all the images which involve these variations [19]. Data augmentation methods are proposed to address this issue by adding the amount and diversity of training samples. Various methods have been proposed focusing on random spatial image augmentation, specifically involving in rotation variation, scale variation, translation, and deformation, etc. [1]. Classical augmentation methods are widely adopted in deep learning techniques.

The main drawback of random data augmentation is low model accuracy. Additionally, it suffers from generating a lot of unavoidable noisy data. Various methods have been proposed to consider data distribution rather than random data augmentation. A search space based data augmentation method has been proposed [20]. It can automatically search for improving data augmentation policies in order to obtain better validation accuracy. In contrast, we leverage WS-DAN [1], which generates augmented data from visual attention features of the image. Peng *et al.* proposed a method for human pose estimation, by introducing an augmentation network whose task is to generate hard data online, thus improving the robustness of models [21]. Nevertheless, their augmentation system is complicated and less accurate compared to the network that we experimented with. Additionally, attention-aware data segmentation is more simple and proven effective in terms of accuracy.

2.2.2 Fine-Grained Visual Categorization

Fine-grained Visual Categorization (FGVC) is a challenging problem in the field of computer vision. Normally, object classification is used for categorize different objects in the image, such as humans, animals, cars, trees, etc. In contrast, fine-grained image classification concentrates more on detecting sub-categories of a given category, like various types of birds, dogs or cars. The purpose of FGVC is to find subtle differences among various categories of a dataset. It presents significant challenges for building a model that generalizes patterns. FGVC is useful in a wide range of applications such as image captioning [22], image generation [23], image search engines, and so on.

Various methods have been developed to differentiate fine-grained categories. Due to the remarkable success of deep learning, most of the recognition works depend on the powerful convolutional deep features. Several methods were proposed to solve large scale real problems [24, 25, 26]. However, it is relatively hard for the basic models to focus on very precise differences of an object's parts without adding special modules [1]. A weakly supervised learning-based approach was adapted to generate class-specific location maps by using pooling methods [27]. Adversarial Complementary Learning (ACoL) [28] is a weakly supervised approach to identify entire objects by training two adversarial complementary classifiers, which aims at locating several parts of objects and detects complementary regions of the same object. However, their method fails to accurately locate the parts of the objects due to having only two complementary regions. On the contrary, our proposed approach depends on attention-guided data augmentation and domain adaptive transfer learning. Our method extracts fine-grained discriminative features and provides a generalization of domain features to achieve state-of-the-art performance in terms of accuracy.

2.2.3 Attention

Attention mechanisms have been getting a lot of popularity in the deep learning area. Visual attention has been already used for FGVC. Xiao *et al.* proposed a two-way attention method (object-level attention and part-level attention) to train domain-specific deep networks [29]. Fu *et al.* proposed an approach that can predict the location of one attention area and extract corresponding features [30]. However, this method can only focus on a local object's parts at the same time. Zheng *et al.* addressed this issue and introduced Multi-Attention CNN (MA-CNN) [31], which can simultaneously focus on multiple body parts. However, selected parts of the object are limited and the number of selected parts is fixed (2 or 4), which might hamper accuracy.

The works mentioned above mostly focus on object localization. In contrast, our research concentrates more on data augmentation with visual attention, which has not been much explored. We use the attention mechanism for data augmentation purposes. Moreover, the benefit of guided attention based data augmentation [1] helps the network to locate object precisely, which helps our trained model learn about closer object details and hence, improve the predictions.

2.2.4 Transfer Learning

The purpose of transfer learning is to improve the performance of a learning algorithm by utilizing knowledge that is acquired from previously solved similar problems. CNNs have been widely used for transfer learning. They are mostly used in the form of pre-trained networks that serve as feature extractors [32, 33].

Considerable amounts of effort have been made to understand transfer learning [34, 35, 36]. Initial weights for a certain network can be obtained from an already-trained network even if the network is used for different tasks [34]. Some prior work has shown some results on transfer learning and domain similarity [7]. Their contribution mostly addresses the effect of image resolution on large scale datasets and choosing different subsets of datasets to boost accuracy. In our work, we show that domain adaptive transfer learning can be useful if we also incorporate visual attention based data augmentation.

Unlike previous works, our proposed technique takes account of domain adaptive transfer learning between the source and target domains. Then, it incorporates the attention-driven approach for data augmentation. Our main goal is to guide the training model to learn relevant features from the source domain and augment data with the visual attention of the target domain. The combination of two processes can be useful to achieve better performance.

2.3 Domain Adaptive Transfer Learning (DATL)

In our research, we explore the way of determining similarity between the source and target domains. Additionally, we describe the attention aware data augmentation technique, WS-DAN in detail. We consider different types of large scale datasets to find out the similarity score between large scale datasets and FGVC datasets. Then, we compute domain similarity score firstly. Based on the domain similarity score we choose large scale datasets for transfer learning and then we perform WS-DAN to evaluate the accuracy.

2.3.1 Domain Similarity

Generally, transfer learning performs better if it has been trained on bigger datasets. Chen *et al.* showed that transfer learning performance increases logarithmically with the number of data [35]. In our work, we observe that using a bigger dataset does not always provide a more accurate result. Yosinski *et al.* [34] mentions that there is some correlation between the transferability of a network from the source task to the target task and the distance between the source and target tasks. Furthermore, they show fine-tuning on a pre-trained network towards a target task can boost performance. Our domain adaptive transfer learning approach is inspired from Cui *et al.* [7] who introduce a method which can calculate domain similarity by the Earth Mover's Distance (EMD) [37]. Furthermore, they show transfer learning can be treated as moving image sets from the source domain S to the target domain T. The domain similarity [7] can be defined

$$d(S,T) = EMD(S,T) = \frac{\sum_{i=1,j=1}^{m,n} f_{i,j} d_{i,j}}{\sum_{i=1,j=1}^{m,n} f_{i,j}} \tag{2.1}$$

where s_i is i-th category in S and t_j is j-th in T, $d_{i,j} = ||g(s_i) - g(t_j)||$, feature extractor $g(.)$ of an image and the optimal flow $f_{i,j}$ computes total work as a EMD minimization problem. Finally, the similarity is calculated as:

$$sim(S,T) = e^{-\gamma d(S,T)} \tag{2.2}$$

where γ is a regularization constant of value 0.01.

Domain similarly score can be calculated between the source and target domain. In our approach, we use large scale datasets as source domains, and target domains are selected from six commonly used FGVC datasets. After calculating the similarity score, we choose top k categories with the highest domain similarity.

2.3.2 Attention Aware Data Augmentation

In our method, we consider using the Weakly Supervised Data Augmentation Network (WS-DAN) [1]. Firstly, we extract features of the image I and feature maps $F \in R^{H \times W \times C}$, where H, W, and C correspond to height, width, and number of channels of a feature layer. Then, we generate attention maps $A \in R^{H \times W \times M}$ from feature maps, where M is the number of attention maps. One more critical component is bi-linear attention pooling, which is used to extract features from part objects. Element-wise multiplication between feature maps and attention maps is computed to get part-feature maps, and then, pooling operation is applied on part-feature maps afterward. Randomly generated data from augmentation is not much efficient. However, attention maps can be handy for data augmentation. This way model can be guided to focus on essential parts of the data and augment those data to the network. With an augmentation map, part's region can be zoomed, and detailed features can be extracted. This process is called attention cropping. Attention maps can represent similar object's part. Attention dropping can be applied to the network to distinguish multiple object's part. Both attention cropping and attention dropping are controlled through a threshold value.

During the training process, no bounding box or keypoints based annotation is available. For each particular training image, attention maps are generated to represent the distinguishable part of object. Attention, guided data augmentation component, is responsible for selecting attention maps efficiently utilizing attention cropping and attention dropping. Bilinear Attention Pooling (BAP) is used to extract features from the object's parts. Element-wise multiplication between the feature

Algorithm 1 Attention Aware Fine-grained Categorization

```
Input: Trained model with WS-DAN and Raw Image I
Output: Classification Accuracy
Calculate coarse-grained probability $p_1 : p_1 = W(I)$ and generate attention maps A
Calculate object map $A_m$ from A and obtain bounding box $B$ from $A_m$
Zoom in the region $B$ as $I_b$
Predict fine-grained probability $p_2 : p_2 = W(I_b)$
Calculate final probability $p = (0.5) * (p_1 + p_2)$
return p
```

maps and attention map are used to generate a part feature matrix. In the last step, the original data, along with attention generated augmented data, are trained as input data.

During the testing process, in the beginning, the object's categories probability and attention maps are produced from input images. Then, the selected part of the object can be enlarged to refine the category's probability. The final prediction is evaluated as the average of those two probabilities. The process of final prediction [1] is presented as Algorithm 1.

The training process is illustrated in Figure 2.1. During training process, no bounding box or keypoints based annotation are available. For each particular training image, attention maps are generated to represent the distinguishable part of object. Attention guided data augmentation component is responsible to select attention maps efficiently utilizing attention cropping and attention dropping. Bilinear Attention Pooling (BAP) is used to extract feature from object's parts. Element-wise multiplication between the feature maps and attention map is used to generate part feature matrix. In the last step, the original data along with attention generated augmented data are trained as input data.

Figure 2.2 shows the illustration of testing process. Firstly, the object's categories probability and attention maps are produced from input images. Then, selected part of the object can be enlarged to refine the categories probability. The final prediction is evaluated as the average of those two probabilities.

2.3.3 Visualization of Augmented Data

We visualize the attention-guided data augmentation in CUB200-2011, Food-101, Flowers-102, Stanford Car, Stanford Dog and FGVC-Aircraft respectively in Figure 2.3-2.8.

2.3.4 Loss Function

The loss function of the network is derived from center loss [38], which has been proposed to tackle face recognition issues. Here, we adopt attention regularization loss [1] for the attention learning process. The idea is to minimize the intra-class variations while keeping the features of inter-class features differentiable. So, penalizing the

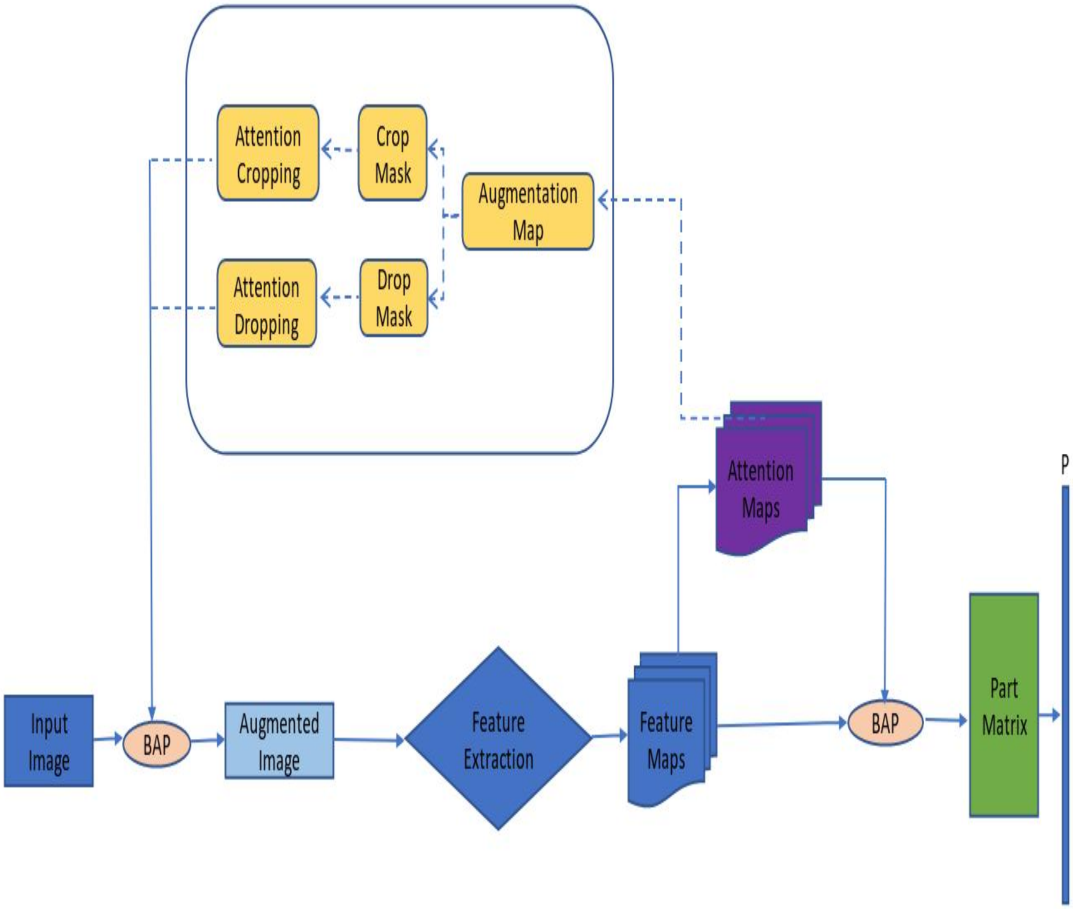

Figure 2.1: Weakly Supervised Data Augmentation Network [1] Training Process.

features variation that belong to same part of object which is important for fine-grained category. The loss function can be defined as:

$$L_A = \sum_{k=1}^{M} ||f_k - c_k||_2^2 \tag{2.3}$$

where M is number of attention maps, f_k is the part feature and c_k is its part's feature center of kth object. c_k can be updated by moving average and initialized as zero, and the update rate is β .

$$c_k \leftarrow c_k + \beta(f_k - c_k) \tag{2.4}$$

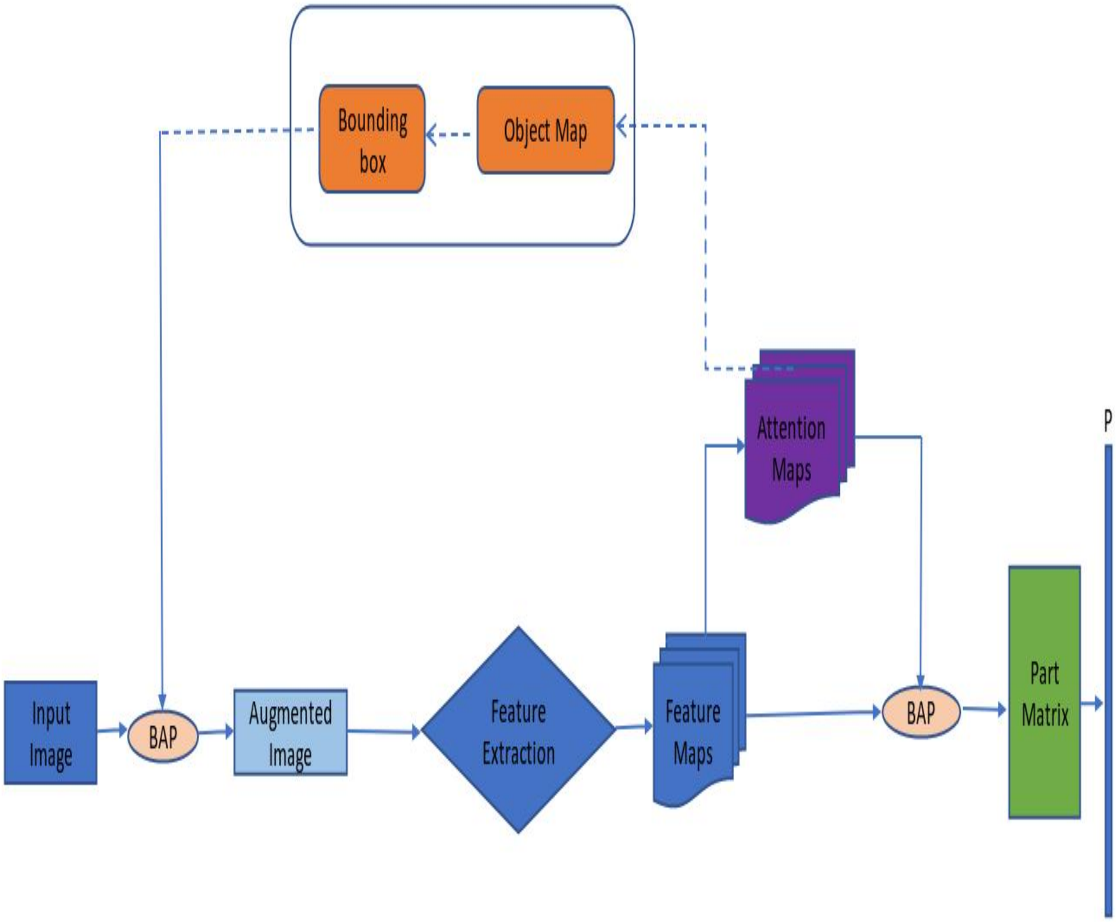

Figure 2.2: Weakly Supervised Data Augmentation Network [1] Testing Process.

2.4 Experiments

In this section, we show comprehensive experiments to verify the effectiveness of our approach. Firstly, we calculate the domain similarity score using EMD [37] to demonstrate the relationship between the source and target domains. Then we compare our model with the state-of-the-art methods on six publicly available fine-grained visual categorization datasets. Following this, we perform additional experiments to demonstrate the effect of image resolution on transfer learning. We compare input

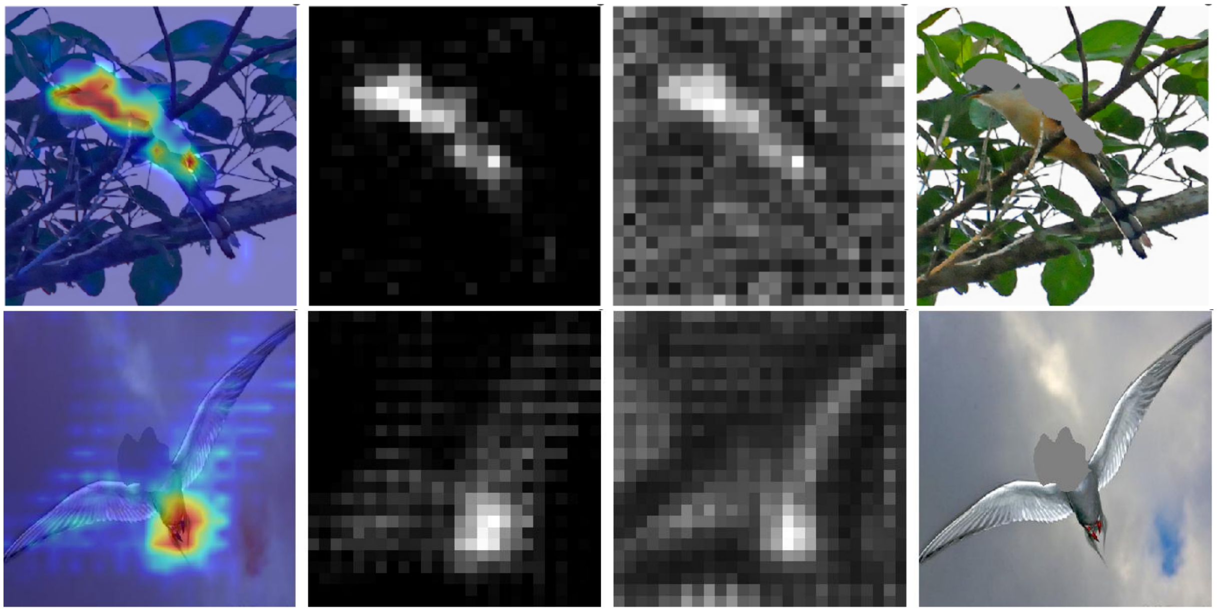

Figure 2.3: Visual attention on image, Attention Maps, Feature Maps, Attention Dropping on CUB200-2011 dataset (left to right respectively) [1].

images in the iNaturalist dataset from 299 × 299 to 448 × 448 to observe the effect in terms of accuracy. We have trained the baseline inceptionV3 model with iNaturalist datasets. Additionally, we combine both iNaturalist and imageNet dataset to make a bigger dataset. We perform detailed experimental studies with different types of large scale datasets and apply the WS-DAN method to observe the impact. The training loss curve and top-1 accuracy curve are presented in Figures 2.9 and 2.10, respectively.

2.4.1 Datasets

We present a detailed overview of the datasets that we use for our experiments.

ImageNet: The ImageNet [12] contains 1.28 million training images and 50 thousand validation images along with 1,000 categories.

iNaturalist(iNat) : The iNat dataset, introduced in 2017 [39], contains more than 665,000 training and around 10000 test images from more than 5000 natural

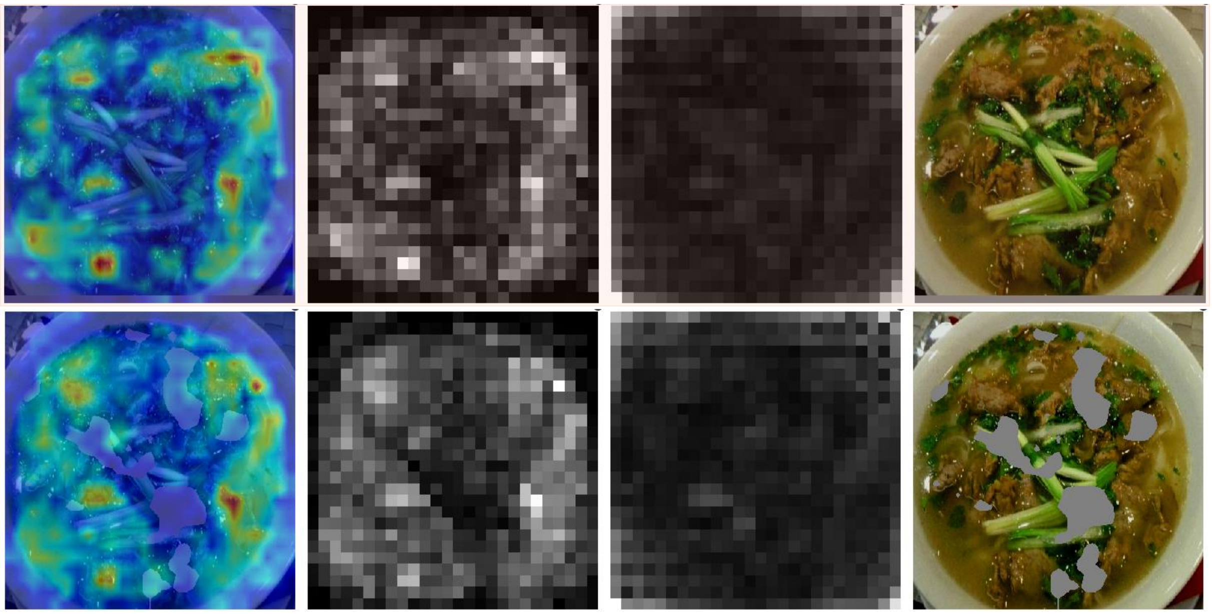

Figure 2.4: Visual attention on image, Attention Maps, Feature Maps, Attention Dropping on Food-101 dataset (left to right respectively) [1].

fine-grained categories. Those categories include different types of mammals, birds, insects, plants, and more. This dataset is quite imbalanced and varies a lot in terms of the number of images per category.

Fine-grained object classification datasets: Table 2.1 summarizes the information of each dataset in detail.

2.4.2 Implementation Details

In our experiment, we used Tensorflow [40] to train all the models on multiple Nvidia Geforce GTX 1080Ti GPUs. The machine has Intel Core-i7-5930k CPU@ 3.50GHz x 12 processors with 64GB of memory. During training, we adopted Inception v3 [26] as the backbone network. We employed WS-DAN [1] technique to perform experiments to demonstrate the effectiveness of transfer learning. For all the datasets, we used Stochastic Gradient Descent (SGD) with a momentum of 0.9, the

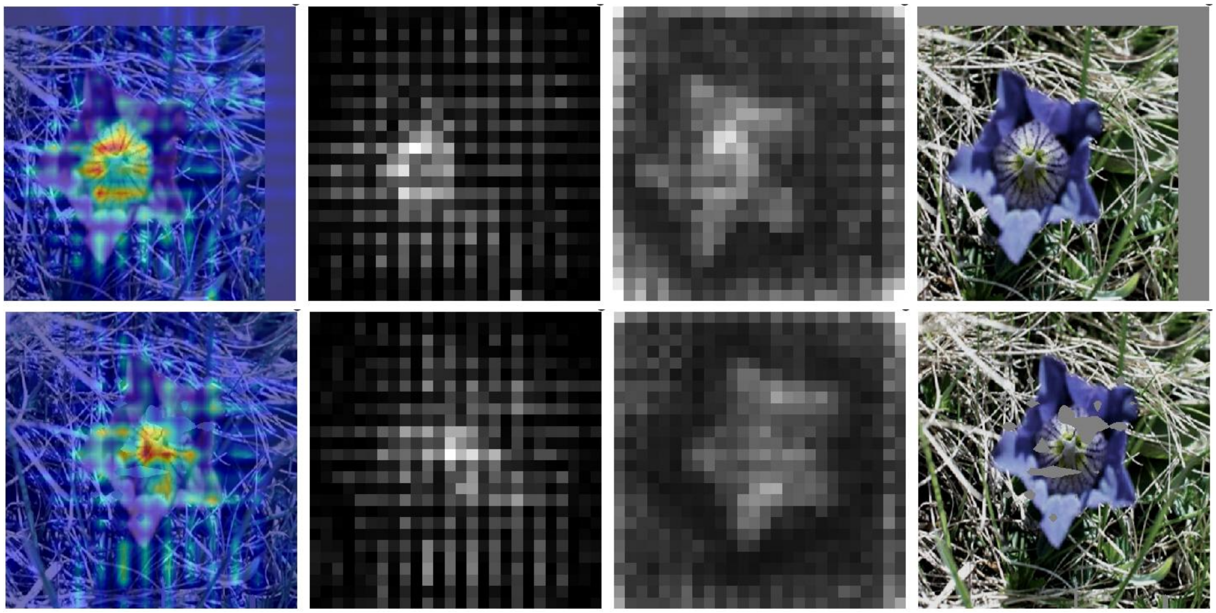

Figure 2.5: Visual attention on image, Attention Maps, Feature Maps, Attention Dropping on Flowers-102 dataset (left to right respectively) [1].

number of epoch 80, mini-batch size 12. The initial learning rate was set to 0.001, with exponential decay of 0.8 after every 2 epochs.

2.5 Results

When training a CNN, input images are often preprocessed to match a specific size. Higher resolution images usually contain essential information and precise details that are important to visual recognition. We compare results on six FGVC datasets with different sizes of image resolution of the iNat dataset. In summary, images with higher resolution yields better accuracy except for the Stanford Dogs dataset. Figure 2.11 represents the effect of transfer learning with various sizes of image resolution on iNat dataset.

In Table 2.3, we present the top-1 accuracy of the target domains on various source domains. These results show the impact of transfer learning from a pre-trained

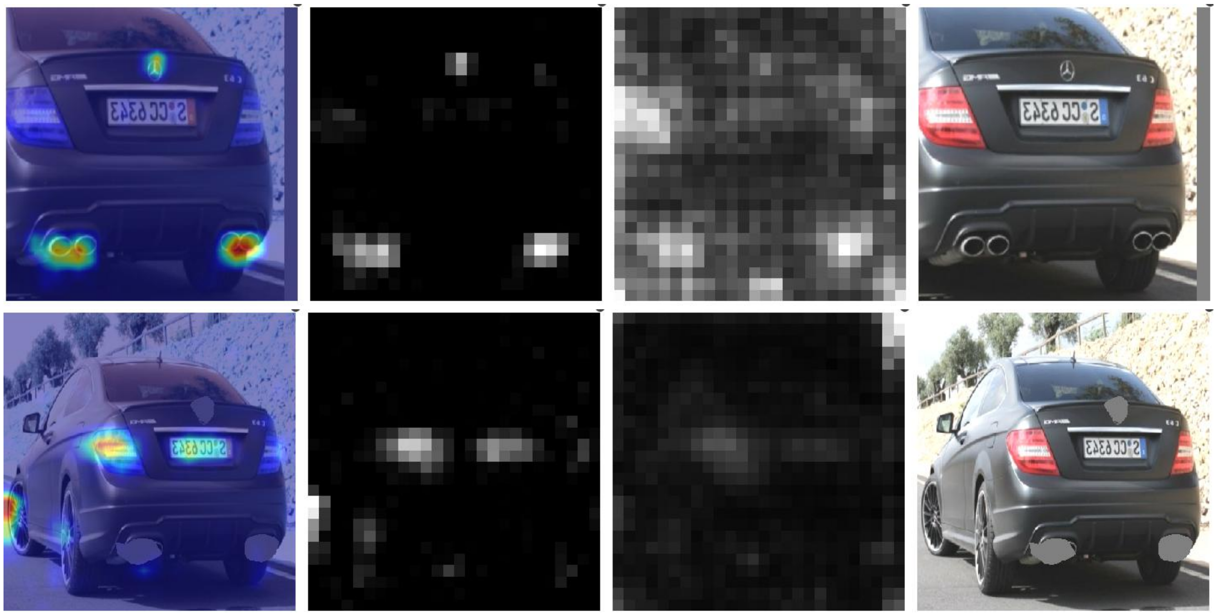

Figure 2.6: Visual attention on image, Attention Maps, Feature Maps, Attention Dropping on Stanford Car dataset (left to right respectively) [1].

model. Large scale datasets are essential for getting improved accuracy when transfer learning is conducted. ImageNet dataset is much larger than iNat dataset; still, it shows worse accuracy in the CUB200-2011 dataset. So, we cannot conclude that using a bigger dataset with transfer learning can always yield better results. Moreover, the domain similarity score also supports this hypothesis. Hence, transfer learning can be effective if the target domain can be trained with similar source domain.

We compare our method with state-of-the-art baselines on six commonly used fine-grained categorization datasets. The summary of the comparison is presented in Table 2.4. In Table 2.2, we show the domain similarity score between the source and various target domains. We visually represent the relationship between the top-1 accuracy and the domain similarity score. We can observe from Figure 2.12 that the domain similarity score positively correlated with transfer learning accuracy between large scale datasets and FGVC datasets. Each marker represents a source domain.

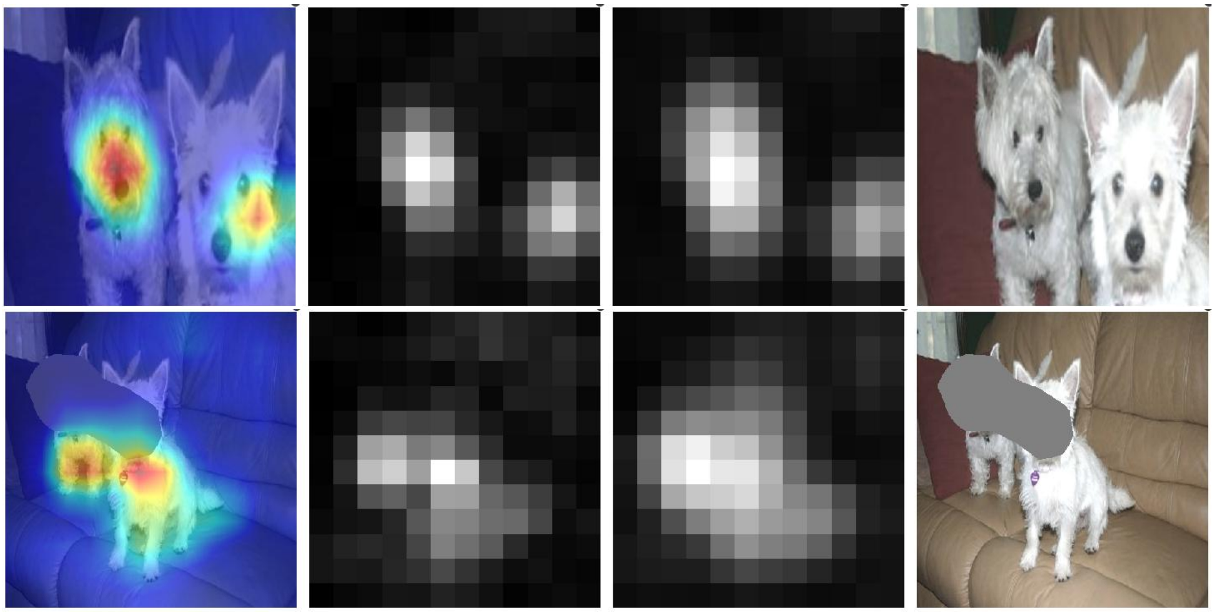

Figure 2.7: Visual attention on image, Attention Maps, Feature Maps, Attention Dropping on Stanford Dog dataset (left to right respectively) [1].

With the right selection of source domain, better transfer learning performance can be achieved. For example, the domain similarity score between iNat and CUB200-2011 is around **0.65**, which is the reason it shows higher accuracy **(91.2)** when iNat is used as pre-training the source domain compared to others. For Flowers-102 dataset, the accuracy is **98.9** with iNat as the source domain which has the highest domain simiarity score **0.54**, among other source domains. Similarly, Stanford Cars, Stanford Dogs and Aircrafts dataset show higher domain similarity score supports better accuracy. Only for the Food101 dataset, the accuracy from transfer learning remains similar while domain similarity changes. We believe this is due to having a large number of training images in Food101. Consequently, the target domain contains enough data and transfer learning is not as useful. We can observe that both ImageNet and iNat are highly biased, achieving dramatically different transfer learning accuracy on target datasets. Intriguingly, when we transfer networks trained

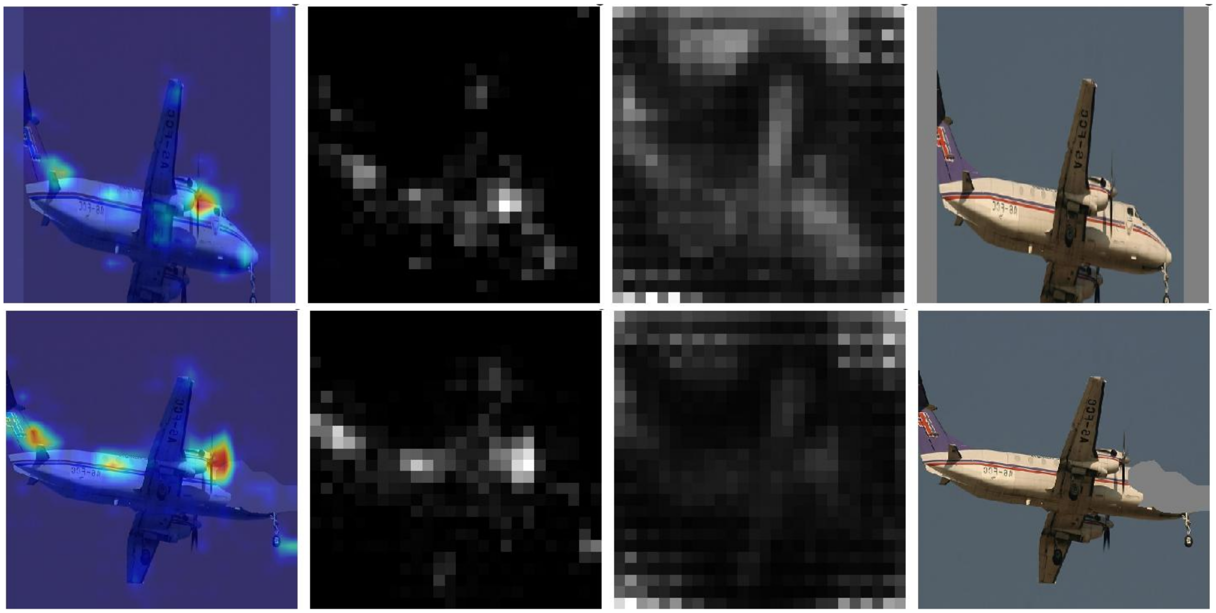

Figure 2.8: Visual attention on image, Attention Maps, Feature Maps, Attention Dropping on FGVC-Aircraft dataset (left to right respectively) [1].

on the combined ImageNet + iNat dataset and perform WS-DAN [1] method over it, we got better results in Food-101 dataset. The resulted accuracy of the combination of ImageNet and iNat, fell in-between ImageNet and iNat pre-trained model. It means that we cannot attain good accuracy on target domains by just using a larger (combined) source domain. Our work demonstrates that a domain similarity score can be useful for identifying which large scale dataset to employ. That way, the model can learn essential features for the target dataset from large source training sets. Furthermore, we can employ attention aware data augmentation techniques to achieve state-of-the-art accuracy on several FGVC datasets.

2.6 Conclusion

In this paper, we describe a simple technique that takes attention mechanism as a data augmentation technique. Attention maps are guided to focus on the ob-

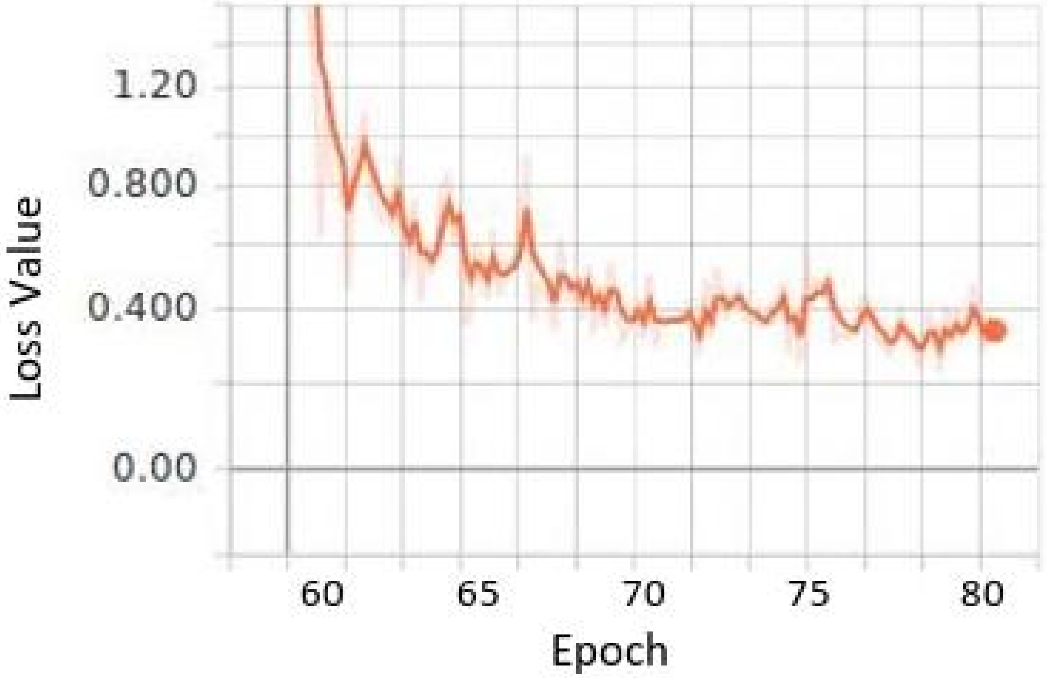

Figure 2.9: Training loss on CUB200-2011 dataset

ject's parts and encourage multiple attention. We demonstrate that domain adaptive transfer learning plays a vital role in boosting performance. Depending on the domain similarity score, we can choose which source datasets to pre-train on to get better accuracy. We show that combining similarity-based selection of source datasets with attention-based augmentation technique can achieve state-of-the-art results in multiple fine-grained visual classification datasets. We also analyze the effect of image resolution on transfer learning between the source and target domains. In future work, we are planning to explore the various factors on transfer learning to boost performance. We like to leverage variational auto encoder and GAN to generate augmented data which can be passed to the model to check the performance. Additionally, we want to compare different types of source datasets and try to control the variability in the number of training images to show the impact.

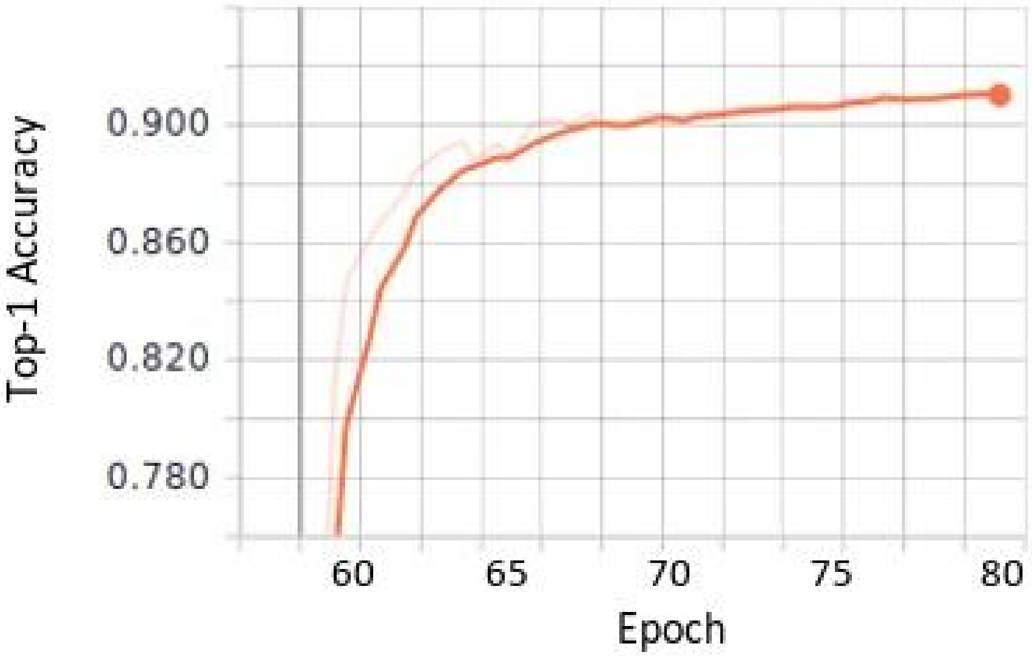

Figure 2.10: Accuracy on CUB200-2011 dataset

Table 2.1: Six commonly used FGVC datasets.

Datasets	Objects	Classes	Training	Test
CUB200-2011	Bird	200	5,994	5,794
FGVC-Aircraft	Aircraft	100	6,667	3,333
Stanford Cars	Car	196	8,144	8,041
Stanford Dogs	Dog	120	12,000	8,580
Flowers-102	Flowers	102	2,040	6,149
Food-101	Food	101	75,750	25,250

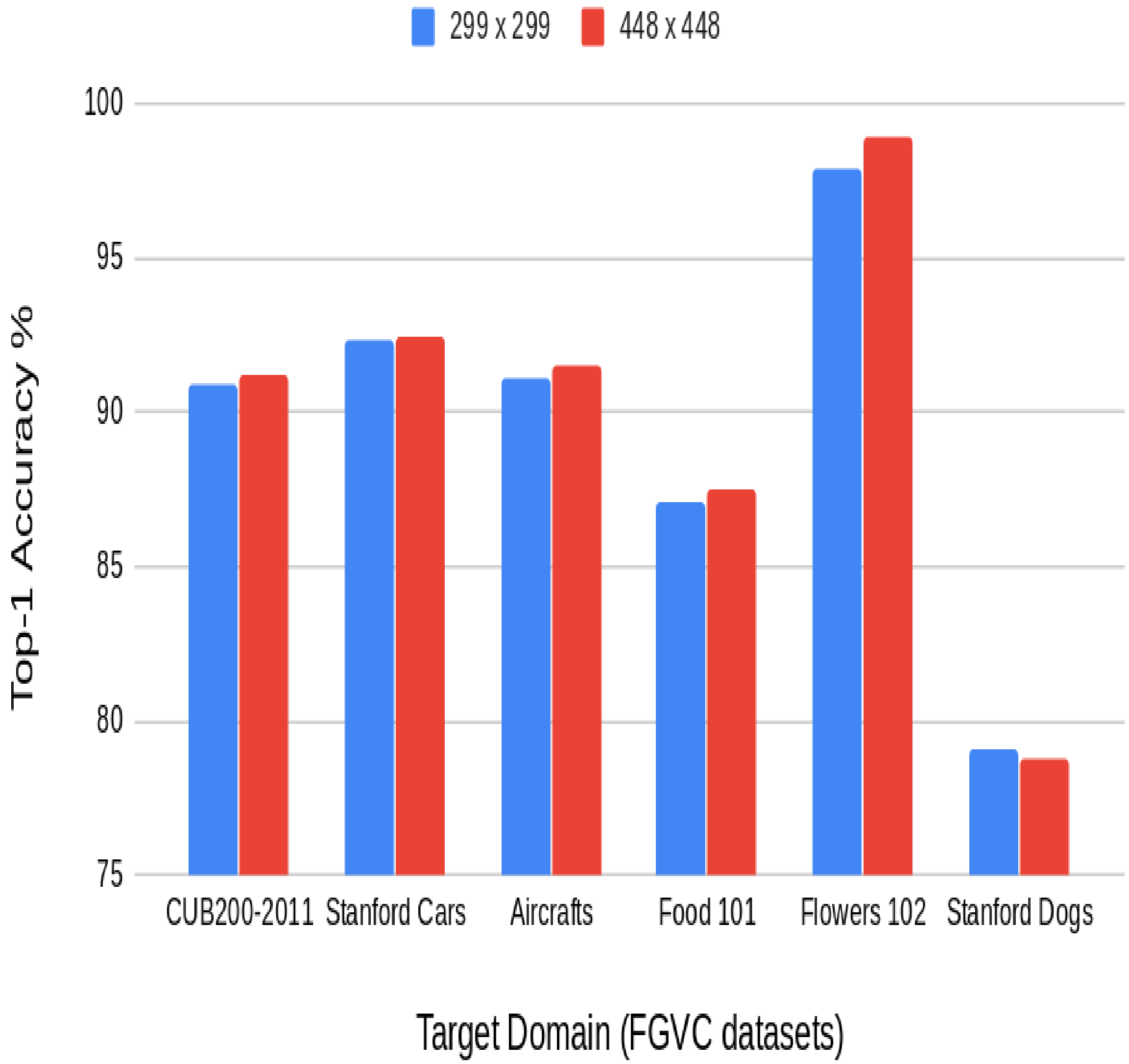

Figure 2.11: Effect of transfer learning with different sizes of image resolution on iNat dataset.

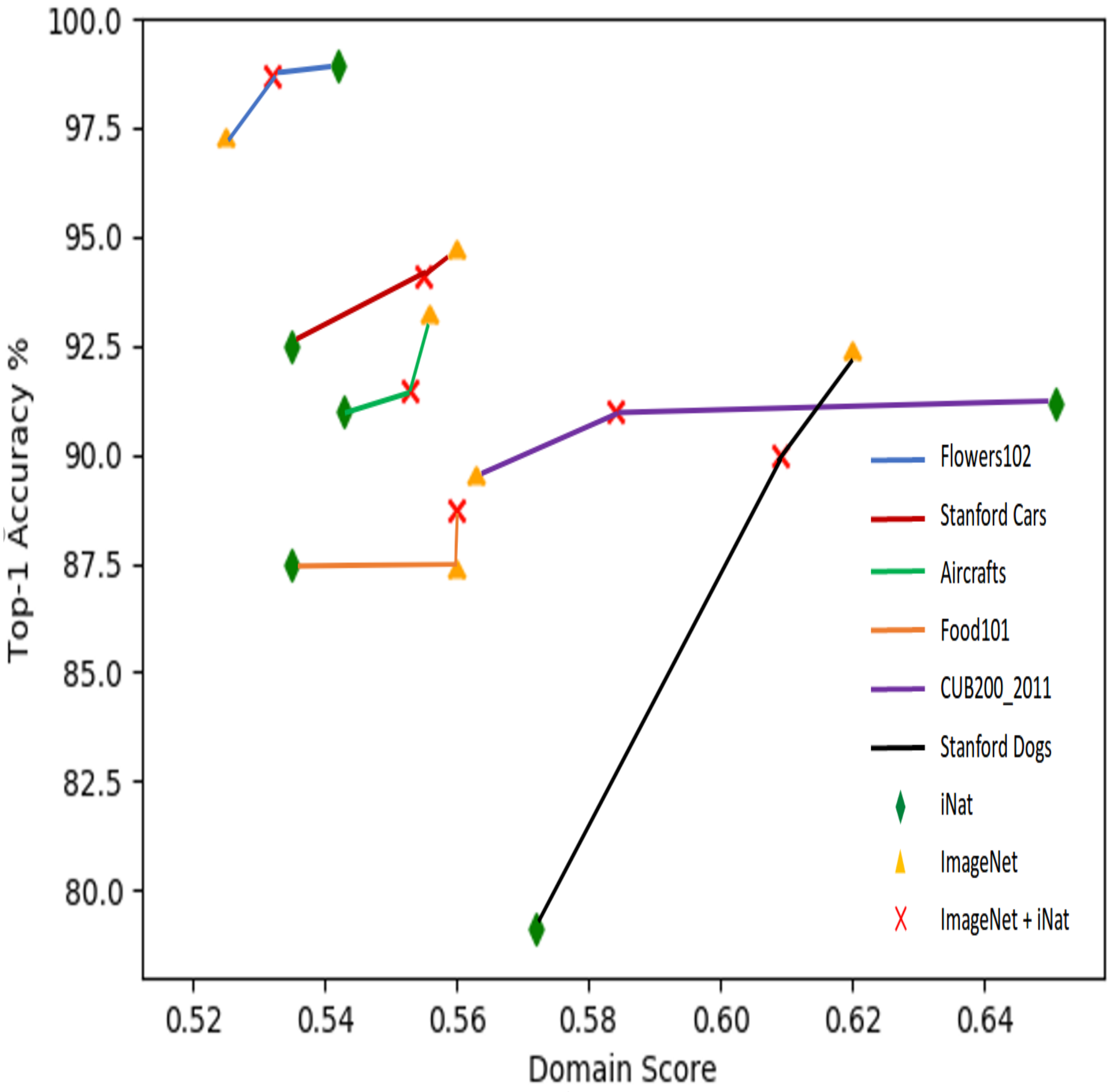

Figure 2.12: Correlation between transfer learning accuracy and domain similarity score between the source and target domain. Each colored line represents a target domain.

Source Domain	CUB200 2011	Stanford Cars	Aircrafts	Food 101	Flowers 102	Stanford Dogs
ImageNet	0.563	0.56	0.556	0.56	0.525	0.619
iNaturalist	0.651	0.535	0.543	0.535	0.542	0.572
ImageNet + iNat	0.584	0.555	0.553	0.56	0.532	0.608

Table 2.2: Comparison on domain similarity score between source datasets and target datasets

Table 2.3: Comparison to different types of FGVC datasets. Each row represents a network pre-trained on source domain for transfer learning and each column represents top-1 image classification accuracy by fine-tuning on the target domain.

Method	CUB200 2011	Stanford Cars	Aircrafts	Food 101	Flowers 102	Stanford Dogs
ImageNet	82.8	91.3	85.5	88.6	96.2	84.2
ImageNet on WS-DAN	89.3	**94.5**	**93.0**	87.2	97.1	**92.2**
iNat on WS-DAN	**91.2**	92.5	91.0	87.5	**98.9**	79.1
ImageNet + iNat on WS-DAN	91.0	94.1	91.5	**88.7**	98.7	90.0

Table 2.4: Comparison in terms of accuracy with existing FGVC methods.

Method	CUB200 2011	Stanford Cars	Aircrafts	Food 101	Flowers 102	Stanford Dogs
Bilinear-CNN [27]	84.1	91.3	84.1	82.4	-	-
DLA [41]	85.1	94.1	92.6	89.7	-	-
RA-CNN [30]	85.4	92.5	-	-	-	87.3
Improved Bilinear-CNN [42]	85.8	92.0	88.5	-	-	-
GP-256 [31]	85.8	92.8	89.8	-	-	-
MA-CNN [30]	86.5	92.8	89.9	-	-	-
DFL-CNN [43]	87.4	93.8	92.0	-	-	-
MPN-COV [44]	88.7	93.3	91.4	-	-	-
Subset B [7]	89.6	93.5	90.7	90.4	-	88.0
WS-DAN [1]	89.4	94.5	93.0	87.2	97.1	92.2
DATL + WS-DAN	**91.2**	94.5	**93.1**	88.7	**98.9**	**92.2**

CHAPTER 3

A Survey on Unsupervised Domain Adaptation for Visual Categorization

3.1 Introduction

Deep learning-based methods have produced remarkable results for many problems in computer vision and machine learning. These methods require a large amount of training and testing data to achieve the expected result. Nevertheless, collecting and annotating datasets for each novel task and domain are extremely costly and time-consuming. One way is to use a pre-trained model and fine-tune it in the target domain. However, fine-tuning demands a large amount of labeled training data. A large amount of labeled data may not be available for many applications.

An overview of domain adaptation is given in Fig 3.1 In the homogeneous setting, the feature spaces between the source and target domains are identical($X^s = X^t$) with the same dimension($d^s = d^t$). Hence, the source and target datasets are generally different in terms of data distributions. In addition, we can classify further the homogeneous DA setting into three cases:

- In the supervised DA, a small amount of labeled target data are present. However, the labeled data are not sufficient enough for tasks.
- In the semi-supervised DA, both limited labeled data and redundant unlabeled data are available in the training state, which allows the networks to learn the structure information of the target domain.
- in the unsupervised DA, no labeled data are available in the training phase.

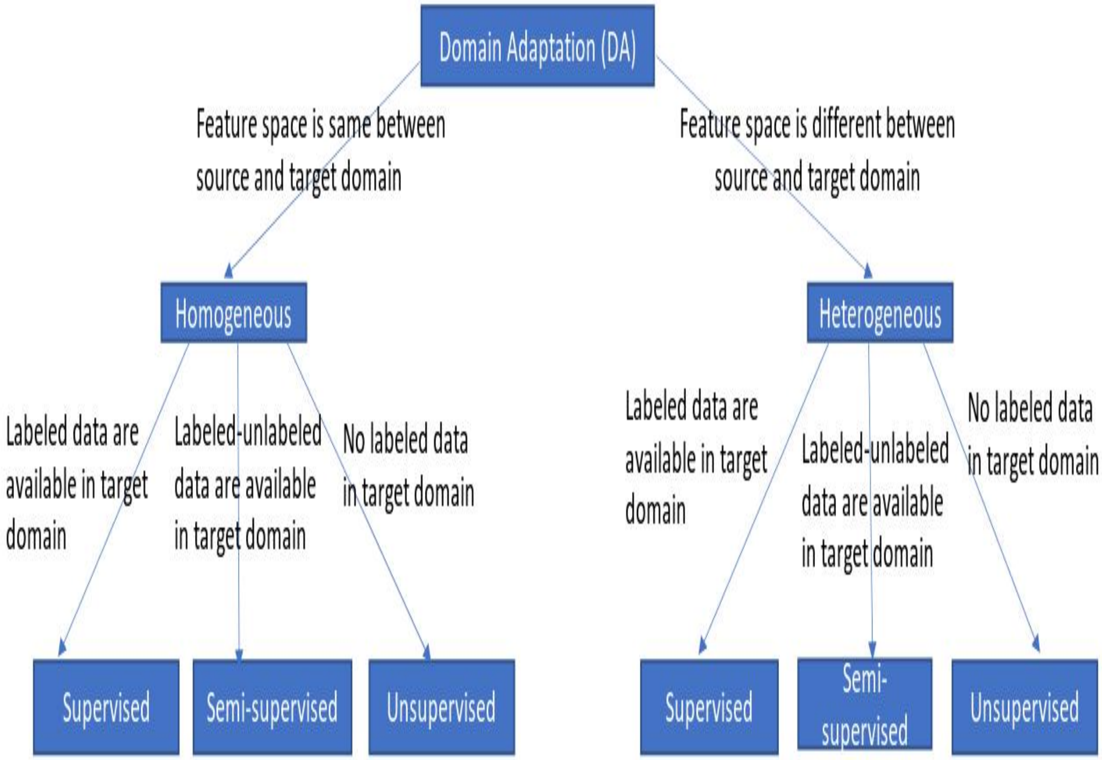
Domain Adaptation (DA)
Feature space is same between source and target domain
Feature space is different between source and target domain
Homogeneous
Heterogeneous
Labeled data are available in target domain
Labeled-unlabeled data are available in target domain
No labeled data in target domain
Labeled data are available in target domain
Labeled-unlabeled data are available in target domain
No labeled data in target domain
Supervised
Semi-supervised
Unsupervised
Supervised
Semi-supervised
Unsupervised

Figure 3.1: An overview of domain adaptation[2].

The feature spaces between the source and target domains are nonequivalent; often, the dimensions may also generally differ. The heterogeneous DA setting can also be categorized into supervised, semi-supervised, and unsupervised DA [2].

This chapter goes over homogeneous unsupervised domain adaptation consisting of single source and single target domain. The main goal of domain adaptation is to reduce domain shift by aligning source and target domains. We can categorize the various alignment methods below.

- **Discrepancy-based Domain Adaptation**: Discrepancy-based domain adaptation refers to fine-tuning the deep network with labeled or unlabeled target data to minimize the domain shift. The most popular method for comparing and diminishing distribution shift are maximum mean discrepancy (MMD) [45, 46, 47, 48], correlation alignment (CORAL) [49, 21], and H divergence and KL divergence [50]

- **Adversarial-based Domain Adaptation**: In this case, a domain discriminator that classifies whether a data is stemmed from the source or target domain is used to confuse domain through an adversarial objective to minimize the distance between the source and target distributions[10, 3, 51].
- **Reconstruction-based Domain Adaptation**: The idea of this approach assumes that the data reconstruction of the source or target samples can help improve the performance of DA. The reconstruction methods are represented using autoencoders [52, 53], and generative adversarial networks (GAN) [54].

Numerous new unsupervised domain adaptation methods have been proposed in recent years, with a growing emphasis on deep learning. We cover existing methods from each category that we mention above.

3.1.1 Divergence-based Domain Adaptation

Divergence-based domain adaptation aims at minimizing some divergence criteria between the source and target domain distributions. The purpose is to achieve domain-invariant feature representation. Maximum Mean Discrepancy (MMD) [45, 3, 55, 56] is a two-sample statistical test of the hypothesis that two distributions are equally based on observed samples from the two distributions. The test is computed from the difference between the mean values of functions on the source and target samples. If the means are different, then the samples are likely not from the same distribution. The samples are mapped into Reproducing Kernel Hilbert Space (RKHS).

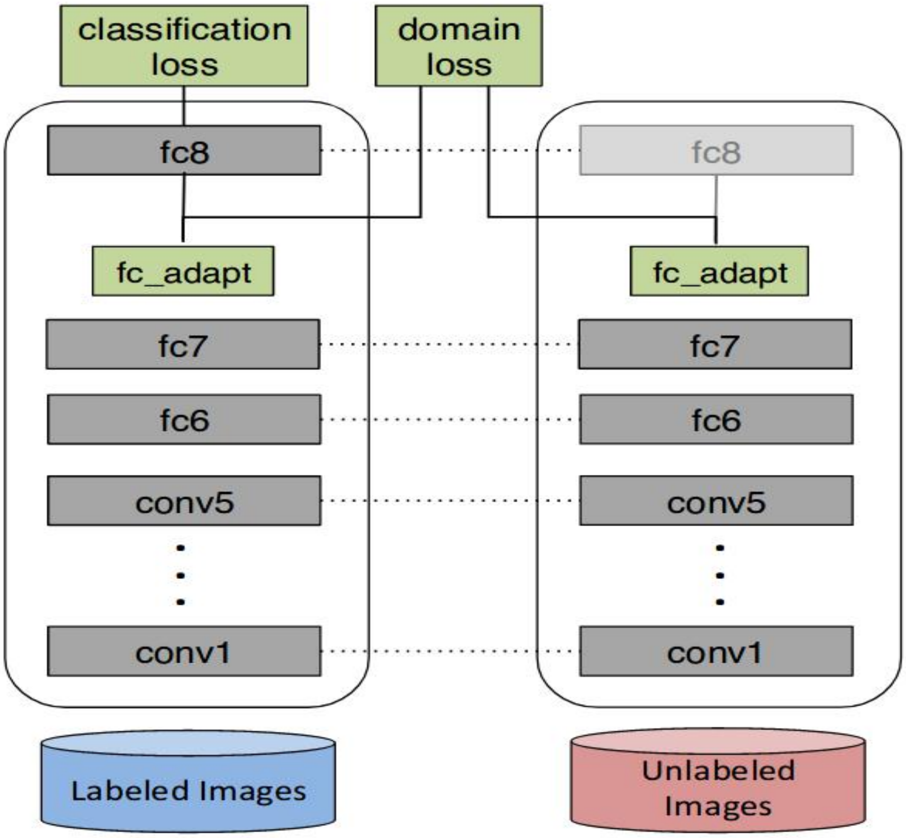

Figure 3.2: The architecture of DDC for learning domain invariant features [3]

A common way is to use the kernel embedding trick and comparing samples using a Gaussian kernel.

In Figure 3.2, Deep Domain Confusion [3] is proposed to adapt MMD in their CNN architecture-based adaptation layer and additional domain confusion loss to learn domain invariant features.

3.1.2 Deep Adaptation Network(DAN)

Recent studies show that the features transferability decreases significantly in the higher layers of a deep network within an increment of domain discrepancy. Practically, the features computed in the higher layers depend a lot on the specific dataset and task. The goal of Deep Adaptation Network (DAN)[4] is to increase the transferability in these task-specific layers by generalizing deep CNN to domain adaptation scenarios. To reach this goal, the hidden representations of all these task-specific layers are embedded to a Reproducing Kernel Hilbert Space (RKHS), where the mean embeddings of various domain distributions can be matched. Because mean embedding matching is affected by the kernel choices, an optimal multi-kernel selection procedure is designed to reduce further the domain discrepancy. The architecture of this method is shown in Figure 3.3.

The starting point of this method is a deep Convolutional Neural Network (CNN) but, just adapting CNN via fine-tuning to the target domain (that has no labels) is complex and could lead to overfitting. Therefore the idea is to model a deep adaptation network (DAN) that can take advantage of both the source domain (with labeled data) and the target domain (with unlabeled data)

The main points that distinguish DAN from previous works are:

Multi-layer adaptation: It is not enough to adapt a single layer to sufficiently reduce the dataset bias between the source and target domain because more than one layer is not transferable. Furthermore, we could link the domain discrepancy that characterizes the marginal and conditional distribution, a critical point for domain adaptation by adapting the representation layers and the classification layer together.

Multi-kernel adaptation: The choice of the kernel is essential; different kernels embed probability distributions in different RKHSs in which different orders of sufficient statistics can be accentuated.

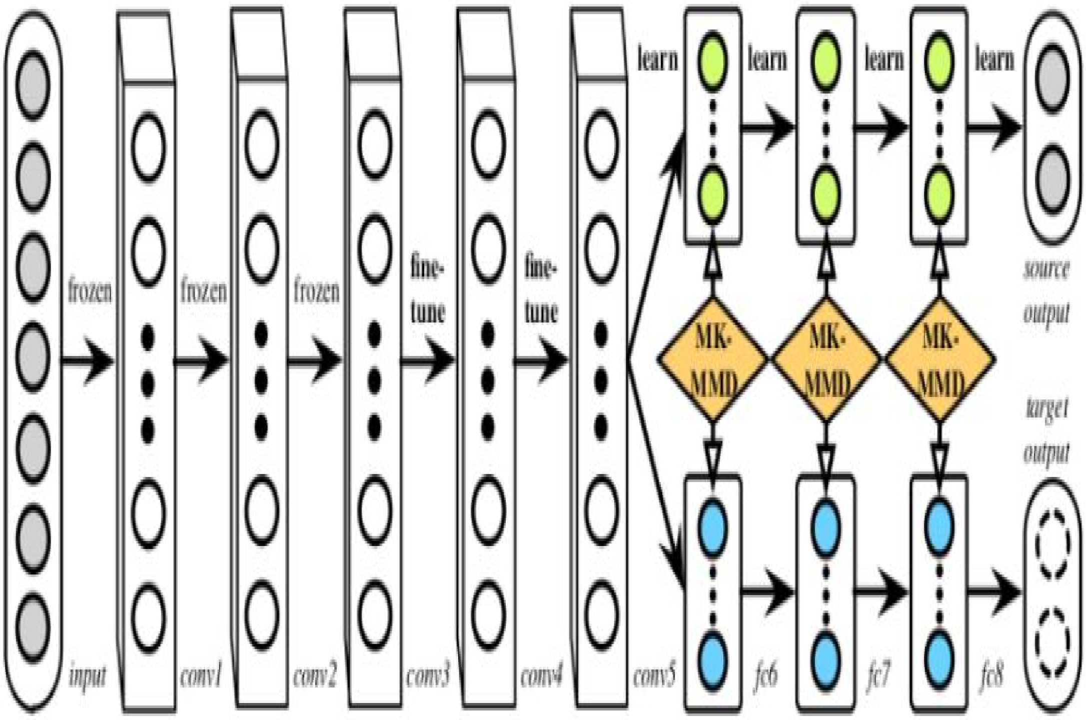

Figure 3.3: The architecture of DAN for learning transferable features [4]

3.1.3 Domain Adversarial Training of Neural Network (DANN)

The core idea of DANN[57] is the implementation of a mapping system between the source and target domain such that the classifier trained on the source works well also for target test data. The algorithm focuses on learning features that are both discriminative and domain-invariant. With this in mind, two classifiers are optimized:

- The label classifier is the standard class labels predictor. The domain classifier whose intention is to distinguish between the source and target domain.
- The optimal feature mapping is found by lowering the loss function of the label classifier and by increasing the loss function of the domain classifier. The

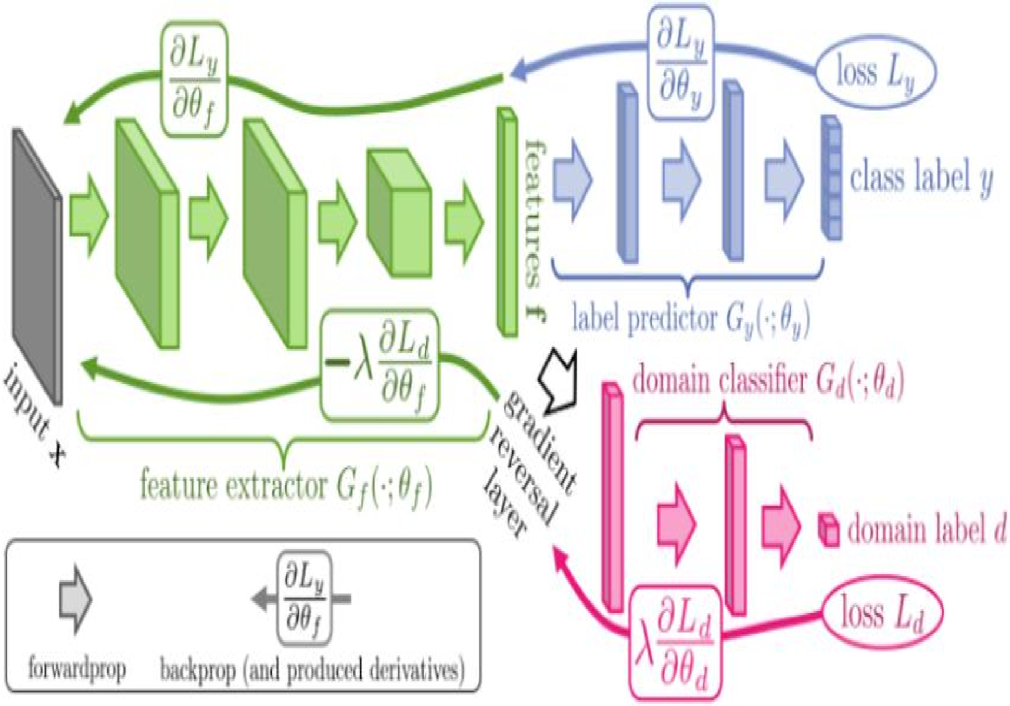
$\frac{\partial L_y}{\partial \theta_f}$
$\frac{\partial L_y}{\partial \theta_y}$
loss L_y
class label y
features $\mathbf{f}$
label predictor $G_y(\cdot;\theta_y)$
input x
$-\lambda \frac{\partial L_d}{\partial \theta_f}$
feature extractor $G_f(\cdot;\theta_f)$
gradient reversal layer
domain classifier $G_d(\cdot;\theta_d)$
domain label d
loss L_d
$\lambda \frac{\partial L_d}{\partial \theta_d}$
forwardprop
backprop (and produced derivatives)

Figure 3.4: The architecture of Domain Adversarial Training of Neural Network with deep features extraction (in green), the deep label predictor (in blue), and the gradient reversal layer (in pink) [4]

maximization then acts adversarially to the domain classifier. The innovation of the DANN algorithm is the addition of Gradient Reversal Layer (GRL) in vanilla CNN.

- Figure 3.4 illustrates the architecture of this method.

3.1.4 Adversarial Discriminative Domain Adaptation (ADDA)

The method[5] is based on a Generative Adversarial Network (GAN)s learning where two networks (a generator and a discriminator) are involved: the generator

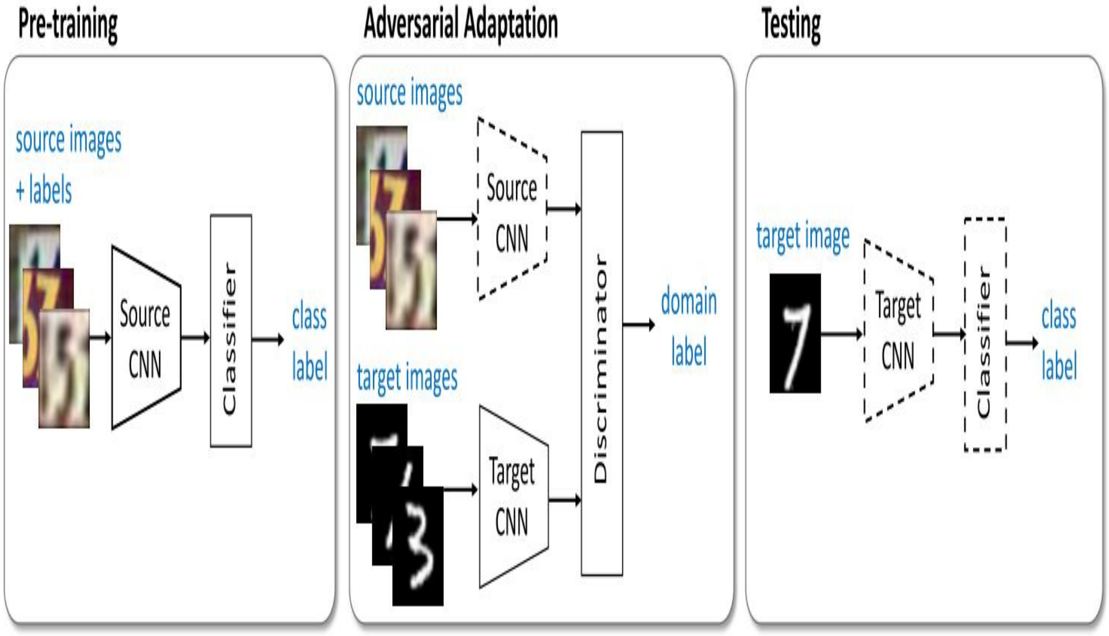

Figure 3.5: The architecture of Adversarial Discriminative Domain Adaptation [5]

yields images, so that confuses the discriminator, which attempts to recognize them from real image examples. This mechanism in domain adaptation is used to ensure that the network cannot distinguish between source and target domain samples. The algorithm implements the following three phases (Figure 3.5).

- Pre-training: a source encoder CNN together with a classifier is trained in a classical way using the labeled source domain.
- Adversarial Adaptation: a target encoder CNN is trained in such a way that a discriminator is not able to recognize the domain label of the examples.
- Testing: the classifier trained in the first phase is used together with the target mapping learned during the second phase to classify the target examples.

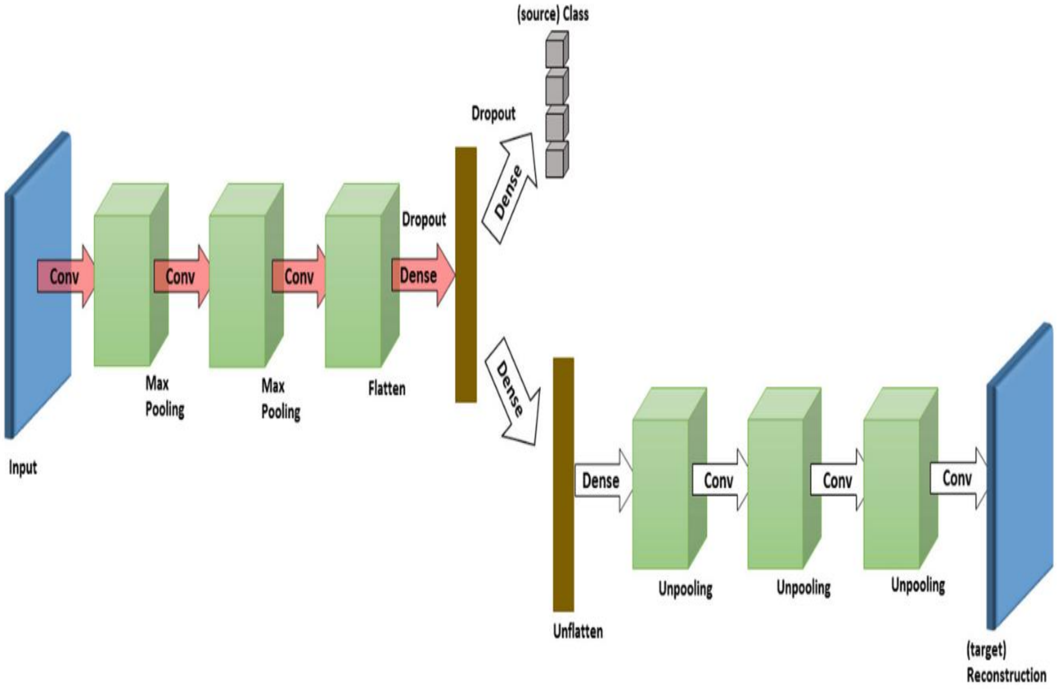

Figure 3.6: The deep reconstruction classification network (DRCN) architecture [4]

3.1.5 Deep Reconstruction-Classification Networks (DRCN)

The deep reconstruction classification network (DRCN) [53] aims to learn a shared encoding representation that provides useful information for cross-domain object recognition. DRCN is a CNN-based architecture that combines two pipelines with a shared encoder. The task for the first pipeline is to classify with source labels, and in the second pipeline, a deconvolutional network optimizes for unsupervised reconstruction with target data. This way, it makes sure that the network learns not only to discriminate correctly but also preserves information about the target data. Additionally, this method mentions that the reconstruction pipeline learns to transform source images into images that resemble the target dataset, which suggests a common representation is learned for both domains. The architecture of this method is depicted in Figure 3.6.

3.2 Discussion

In this chapter, we provide an overview of unsupervised domain adaptation on visual tasks, mainly using deep learning methods. Table 3.1 summarizes deep neural network-based domain adaptation methods where each one shows different loss functions and different types of adaptation techniques.

Table 3.1: Comparison of deep learning based domain adaptation methods on different settings

Method	Discrepancy Loss	Adversarial Loss	Generator
CAN [58]	Contrastive Domain Discrepancy (CDD)		
SimNet[59]		Feature	
Rozant. et al. [45]	Maximum Mean Discrepancy (MMD)		
MCD [51]	Maximum Classifier Discrepancy (MCD)		
SimGAN [60]	Pixel	GAN	
DRCN [53]	Cycle Consistency		
DANN [10, 57]		Feature	
DAN [4]	MK-MMD		
PixelDA [52]		Pixel	GAN
CoGAN [61]			GAN
MADA [62]		Feature	
Deep CORAL [49]	CORAL		
Domain Mixup [63]		Feature + Pixel	
ADDA [5]		Feature	

CHAPTER 4

Adaptive Feature Norm for Unsupervised Sub-Domain Adaptation

4.1 Introduction

In human learning, typically a child can learn new concepts without the need for millions of examples that are pointed out individually. Once a child has seen one cat, she or can recognize other cats to some extent, and gets better at recognizing cats without seeing cats a thousand times. A main motivation of domain adaptation is to train a system which can learn how to perform a similar task, e.g., recognizing objects, like a child.

Deep Neural Networks have shown remarkable performance in various domains in the field of computer vision. To achieve good performance, they typically require a vast amount of labeled data. Training larger and deeper networks is complicated if the size of a dataset is small. Additionally, collecting well-annotated data is costly and time-consuming. A popular way to regularize these networks is to simply use a pre-trained model trained on a different dataset and use the model for the target dataset. However, if the data distribution between source and target domains is different, it may lead to adverse effects and hamper the generalization ability of the models [64]. Unsupervised Domain Adaptation (UDA) focuses on transferring knowledge from a labeled source domain to an unlabeled target domain, and a large amount of research tries to achieve this by exploring domain-invariant representations to bridge the gap. Traditional machine-learning paradigms, like supervised learning, tend to train models to predict the outcome for unseen data. These models do not necessarily optimize performance if there is enough difference between the test and training data

[65]. According to Tzeng *et al.*[3], while generically trained deep networks have a reduced dataset bias, there still exists a domain shift between different datasets, and it is required to adapt the features appropriately. [66] suggests that a fair domain adaptation method should be based on features that are near similar for the source and target domains while reducing the prediction error in the source domain as much as possible. However, domain adaptation can have a domain shift problem. For example, the target domain may contain images from different imaging device (e.g. webcam vs. dslr camera), resulting in different styles in photos. This means the object recognition model trained from the source domain requires to be adapted to the target domain. Therefore, to reduce the domain shift problem, the two domains marginal distributions need to be as similar as possible. The primary goal of UDA is to learn domain-invariant feature representations that can reduce the domain shift. According to existing studies, domain-invariant representations can be captured through several methods, e.g., Maximum Mean Discrepancy [67, 4, 68], divergence-based methods [66, 51], correlation distance [49], etc. Addtionally, several adversarial based methods have been applied [10, 5, 69, 70, 71] to minimize an approximate domain discrepancy.

Recent studies have shown that, compared to shallow networks, deep networks can learn more transferable features for domain adaptation by extracting domain-invariant features [34, 4, 72, 49]. The main observation from the previous domain adaptation methods is that the domain classifier should be confused maximally so that the source classifier treats the samples from the target domain in a similar fashion. Additionally, most successful methods have come up with such ways that can make the domain classifier more confused. Most of the previous domain adaptation methods consider aligning the source and target distributions globally. We adapt a subdomain based approach to learn the domain transfer. A subdomain consists of samples within the same class. This method will lead to a scenario where all the data

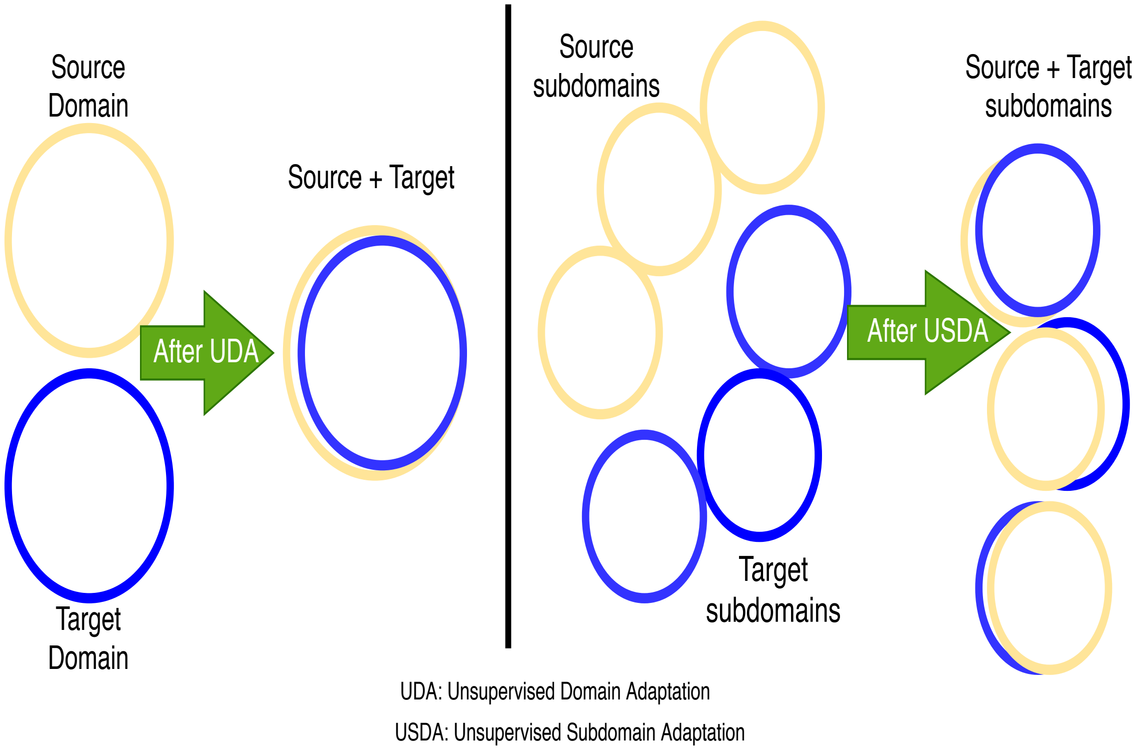
Source Domain
After UDA
Source + Target
Target Domain
Source subdomains
Target subdomains
After USDA
Source + Target subdomains
UDA: Unsupervised Domain Adaptation
USDA: Unsupervised Subdomain Adaptation

Figure 4.1: Domain Adaptation Vs. Subdomain Adaptation

from different domains will be confused, and discriminative structures can be mixed up[48]. The main advantage of the subdomains over domains is the local domain shift instead of the global domain shift. Because of the local domain shift, the learners precisely may align the distribution of relevant subdomains within the same category in the source and target domains. An illustrative example of the difference between Domain Adaptation and Subdomain adaptation is depicted in Figure 4.1. After global domain adaptation, the resulting distributions of the two domains are quite similar, but the data in different subdomains are too adjacent to be correctly classified. The distributions of relevant subdomains can be aligned properly, hence the nuances of the information can be exploited for domain adaptation.

According to [73], larger norms enable more informative trasferability. Recent studies on the compression technique [74] support the above claim and suggest smaller norms contain less information during the inference. Inspired from the two studies as mentioned above, we incorporate the step-wise adaptive feature norm approach in subdomain space.

Xu *et al.*[73] demonstrates that progressively adapting the feature norms of two domains to a broad range of values can boost domain transfer. We present the local maximum mean discrepancy based method with adaptive progressively feature norm on subdomain space. For effective UDA, our goal is to endorse positive transfer and circumvent negative transfer.

In summary, the main contributions of our work are:

1. We propose an innovative stepwise adaptive feature norm-based approach for unsupervised subdomain adaptation. This approach employs to learn task-specific features in a progressive manner, which assists in aligning relevant subdomains in unsupervised scenarios.
2. We demonstrate a local MMD [48] based method with stepwise adaptive feature norm to achieve state-of-the-art results on Office-31, Office-Home, and Image-CLEF datasets.
3. We comapare our results with both adversarial and non-adversarial methods to show the efficacy of our work.

4.2 Related Work

Domain adaptation problem has been widely studied in the computer vision research community. Various methods have been employed to generalize the model across different domains by mitigating the domain shift problem. This section will discuss the relevant work in domain adaptation, subdomain adaptation, and maximum mean discrepancy.

4.2.1 Domain Adaptation

Domain adaptation can be a way to mitigate domain shift issues and reduce the effort of recollecting and retraining a model by transferring knowledge between tasks and domains. Domain adaptation can be defined as the task of training a model on labeled data from a source domain while minimizing test error on a target domain, where no labels for the target domain are available at training time. Several types of methods have been employed for unsupervised domain adaptation. Discrepancy based methods explore domain-invariant structures by reducing some specific statistic distances between the two domains. Maximum Mean Discrepancy (MMD) [75] has been adopted in many approaches [67, 4] for domain adaptation. It enables the model to learn transferable features by reducing the MMD of their kernel embedding. Some other methods extended MMD [46, 47] to measure the source and target data's joint distributions. In our case, we consider local MMD measures discrepancy of relevant subdomains between source and target domains. Adversarial DAs [10, 5, 71] are widely applied in this field. They involve a sub-network as the domain discriminator to distinguish features of alternate domains, whereas learners try to generate features that confuse the domain classifier.

4.2.2 Subdomain Adaptation

A significant amount of research for subdomain adaptation has been published recently. Multiadversarial domain adaptation (MADA) captures the multimode structures to enable fine-grained alignment of various data distributions [62]. CDAN [72] captures the adversarial domain adaptation on discriminative information to enable alignment of multimodal distributions. Moving the semantic transfer network (MSTN) [70] captures the semantic representation for unlabeled target samples by aligning the source and target centroid. Another method [76] creates multiple diverse feature spaces and aligns the source and target distributions in each of them separately while encouraging that alignments agree with each other with regard to the class predictions on the unlabeled target samples. All these methods have adopted adversarial loss. Compared to our work, we have adopted a discrepancy based strategy with stepwise adaptive feature norm approach which is more straightforward and can perform better than these previous methods.

4.2.3 Maximum Mean Discrepancy

Among discrepancy based methods, MMD is one of the most popular metrics of training for domain invariant features. In Deep Adaptation Network (DAN) architecture [4], the authors train the first layers of the model commonly with the source and target domains; after that, they train individual task-specific layers while minimizing MMD between layers. Additionally, MMD has been extended by [46, 47]. However, most previous work considers global MMD measures to reduce discrepancies between the source and target samples. Our work is based on local MMD [48], which measures the discrepancy in relevant subdomains between the source and target domains.

Compared to the previous technique, we use a non-adversarial based subdomain adaptation method and incorporate adaptive feature norms within the subdomains to perform domain transfer. So, instead of just relying on a particular discrepancy metric, we take an additional approach as an adaptive feature norm. In our framework (Figure 4.2), we have shown that a progressive feature-norm-based loss function in a shared subdomain space can boost the domain adaptation performance.

4.3 Method

In unsupervised domain adaptation, we are given a source domain $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ of n_s labeled examples and a target domain $D_t = \{x_j^t\}_{j=1}^{n_t}$ of unlabeled examples. The

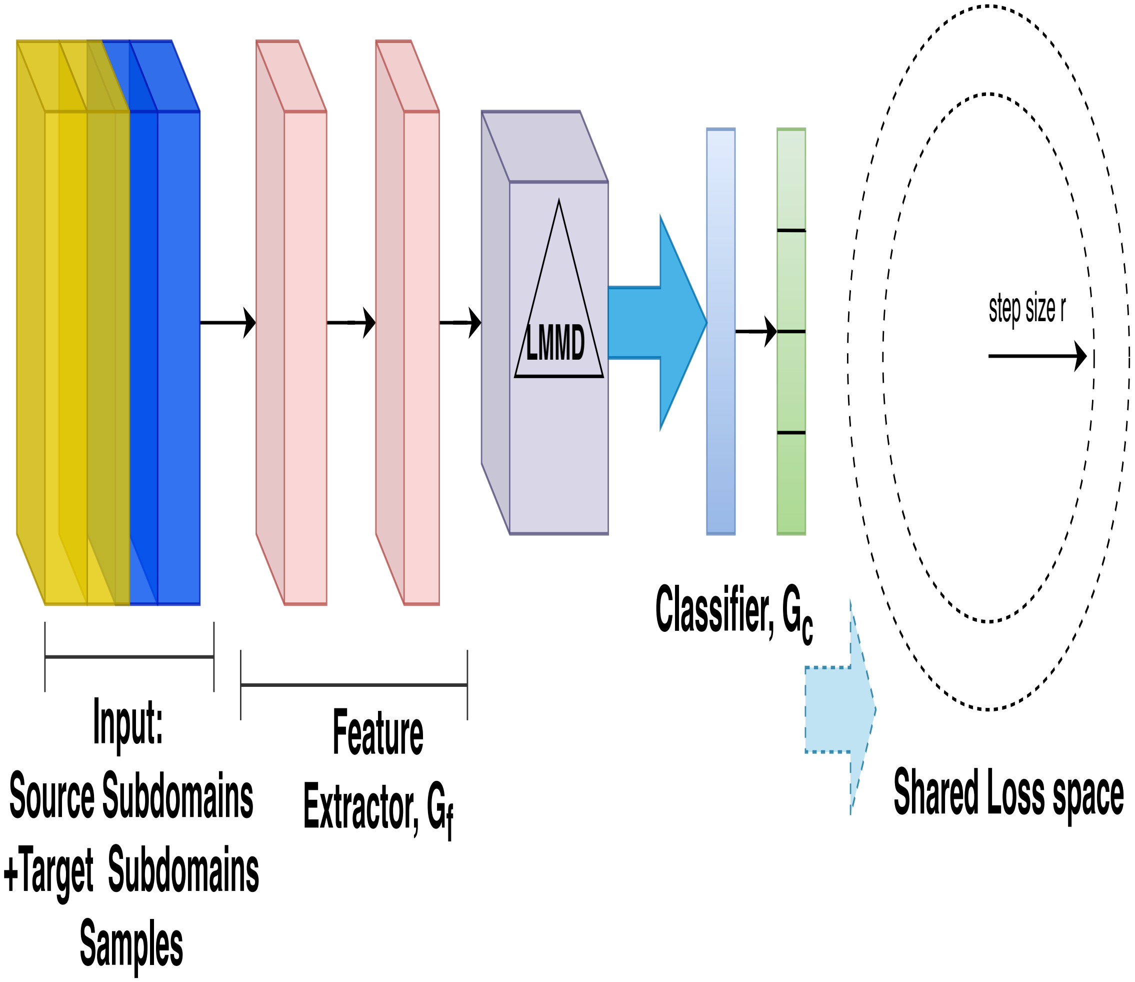

Figure 4.2: Architecture of adaptive feature norm on unsupervised subdomain adapatation.

source domain and target domain are sampled from joint distributions $P(X^s, Y^s)$ and $Q(X^t, Y^t)$ respectively, where $P \neq Q$. The goal of our method is to develop a deep network architecture, that contains transfer features $f = G_f(x)$ and adaptive classifier $y = G_c(f)$. This model will minimize the shift in joint distribution across relevant subdomains and learns transferable representations simultaneously.

The formal representation for unsupervised domain adaptation is as follows.

$$\min_f \frac{1}{n_s} \sum_{i=1}^{n_s} J(f(X_i^s), y_i^s) + \lambda \hat{d}(p, q) \tag{4.1}$$

where $J()$ is the cross-entropy loss function (classification loss) and $\hat{d}()$ is domain adaptation loss. $\lambda > 0$ is the trade-off parameter of the domain adaptation loss and the classification loss.

This representation covers the global source and target domain without taking into account the relevant information between subdomains within the same category between the source and target domains. Nevertheless, the global alignment may not capture the nuances among subdomains. This may lead to domain shift issue as well. The subdomain information can exploit the relationship between different domains. The formal representation of the loss of subdomain adaptation can be

$$\min_f \frac{1}{n_s} \sum_{i=1}^{n_s} J(f(X_i^s), y_i^s) + \lambda \mathbf{E}_c[\hat{d}(p^{(c)}, q^{(c)}]) \tag{4.2}$$

where $\mathbf{E}_c[.]$ is a mathematical expectation of the class.

4.3.1 Local Maximum Mean Discrepancy

We have used local MMD as the baseline architecture. It was proposed by [48] to align distributions of the relevant subdomains.

$$d_H(p, q) = \mathbf{E}_c ||\mathbf{E}_{p^{(c)}}[\phi(x^s)] - \mathbf{E}_{q^{(c)}}[\phi(x^t)]||_H^2 \tag{4.3}$$

where x^s and x^t are the instances in D_s and D_t, and $p^{(c)}$ and $q^{(c)}$ are the distributions of $D_s^{(c)}$ and $D_t^{(c)}$, respectively. The equation (3) can measure class by class difference of the relevant subdomains. Additionally, this can be used to align the subdomains within the target domain with those in the source domain. Since we have an assumption that each sample belongs to each class according to weight w^c, we use an unbiased estimator of equation (3) as

$$d_H(p, q) = \frac{1}{C} \sum_{c=1}^{C} || \sum_{x_i^s \in D_s} w_i^{sc} \phi(x_i^s) - \sum_{x_j^t \in D_t} w_j^{tc} \phi(x_j^t)||_H^2 \tag{4.4}$$

where w_i^{sc} and w_j^{tc} represent the weight of x_i^s and x_j^t belonging to class c, respectively. The sum of weights are both equal to one. We can formulate w_i^c for the sample x_i as

$$w_i^c = \frac{y_{ic}}{\sum_{(x_j, y_j \in D} y_{jc}} \tag{4.5}$$

where y_{ic} is the cth entry of vector y_i. For source domain, we use the ground truth y_i^s as a one-hot vector to calculate w_i^c for each sample. But, for target domain, we use the probability of assigning x_i^t to each of the classes. we can not use the formula of equation (4) directly. The output of the deep neural network is a probability distribution. We use that probability distribution which characterizes the probability of assigning samples to the classes for each target sample. Then, we can calculate w_j^{tc}. Finally, we can calculate equation (4).

4.3.2 Architecture

Standard domain adaptation considers two domains to share a similar label space. In our framework, the input consists of subdomains from the source and target domains. We have a backbone network G_f, which denotes the feature extraction module. Classifier G_c is the task-specific classifier. We apply the feature norm adaptation along with the local MMD based method to optimize the source classification loss during each iteration. In each iteration, each individual sample's feature norm is getting added a small but progressive step size of r. This way, if any target samples are far way from the small norm region, after the domain adaptation step, it could be classified correctly in an automatic manner. Figure 4.2 demonstrates the architecture of our approach.

4.3.3 Adaptive Feature Norm Loss

One of the major bottlenecks that we observe is smaller feature norm of the source and target samples that can lead to poor transfer gains[73]. Inspired from them, we extend the idea into subdomain spaces. We keep a parameter r, which progressively modifies the mean feature norm in each iteration. Instead of having a fixed feature norm, we consider a moving parameter which changes the mean feature norm. This method has been unexplored for the subdomain adaptation case. This loss value impacts the target samples to be correctly classified without additional supervision. This variant impacts positively towards learning task-specific features in a continuous manner. We propose

$$\begin{aligned} \hat{d}_H(p,q) &= \mathbf{E}_c||\mathbf{E}_{p(c)}[\phi(x^s)] - \mathbf{E}_{q(c)}[\phi(x^t)]||^2_H \\ &+\mathbf{E}_c||\mathbf{E}_{p(c)}[(h(x_i;\theta_0)+\Delta r), h(x_i;\theta)]|| \end{aligned} \tag{4.6}$$

where $h(x) = ||.||_2 \cdot G_f \cdot G_c(x)$, where θ_0 and θ are model parameters of last and current iterations. The effectiveness of this model parameter enables the optimization process fetching more informative features with larger norms.

4.4 Experiment

We evaluate our technique on three popular object recognition datasets, including Office-31, Office-Home, and ImageCLEF-DA. The code will be published in future.

4.4.1 Dataset

We present a detailed overview of the datasets that we use for our experiments.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet[25]	68.4 ± 0.5	96.7 ± 0.5	99.3 ± 0.1	68.9 ± 0.2	62.5 ± 0.3	60.7 ± 0.3	76.1
DDC[3]	75.8 ± 0.2	95.0 ± 0.2	98.2 ± 0.1	77.5 ± 0.3	67.4 ± 0.3	64.0 ± 0.5	79.7
D-CORAL[49]	77.7 ± 0.3	97.6 ± 0.2	99.5 ± 0.1	81.1 ± 0.4	64.6 ± 0.3	64.0 ± 0.4	80.8
DAN[4]	83.8 ± 0.4	96.8 ± 0.2	99.5 ± 0.1	78.4 ± 0.2	66.7 ± 0.3	62.7 ± 0.2	81.3
DANN[57]	82.0 ± 0.4	96.8 ± 0.2	99.1 ± 0.1	79.7 ± 0.4	68.2 ± 0.4	67.4 ± 0.5	82.2
ADDA[5]	86.2 ± 0.5	96.2 ± 0.3	98.4 ± 0.3	77.8 ± 0.3	69.5 ± 0.4	68.9 ± 0.5	82.9
JAN[47]	85.4 ± 0.3	97.4 ± 0.2	99.8 ± 0.2	84.7 ± 0.3	68.6 ± 0.3	70.0 ± 0.4	84.3
MADA[62]	90.0 ± 0.1	97.4 ± 0.1	99.6 ± 0.1	87.8 ± 0.2	70.3 ± 0.3	66.4 ± 0.3	85.2
CDAN[72]	93.1 ± 0.2	98.2 ± 0.2	100 ± 0	89.8 ± 0.3	70.1 ± 0.4	68.0 ± 0.4	86.6
iCAN[77]	92.5 ± 0.2	98.8 ± 0.1	100 ± 0	90.1 ± 0.1	72.1 ± 0.2	69.9 ± 0.1	87.2
CDAN + E[72]	94.1 ± 0.1	98.6 ± 0.1	100 ± 0	92.9 ± 0.2	73.5 ± 0.5	69.3 ± 0.3	87.7
DSAN[48]	93.4 ± 0.2	98.3 ± 0.1	100 ± 0	90.2 ± 0.7	73.5 ± 0.5	74.8 ± 0.4	88.2
Ours	93.2 ± 0.2	98.7 ± 0.2	100 ± 0	90.1 ± 0.2	75.1 ± 0.3	72.8 ± 0.4	**88.5**

Table 4.1: Accuracy Comparison of Unsupervised Domain Adaptation on Office-31 Dataset

Method	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R
ResNet[25]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4
DAN[4]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7
DANN[57]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5
JAN[47]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3
CDAN[72]	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9
CDAN+E[72]	50.7	70.8	76.0	57.6	70.0	70.0	57.4	50.9	77.3
DSAN[48]	54.4	70.8	75.4	60.4	67.8	68.0	62.6	55.9	78.5
Ours	55.0	71.0	75.3	61.1	69.4	68.0	61.4	55	78

Method	R→A	R→C	R→P	Avg
ResNet[25]	53.9	41.2	59.9	46.1
DAN[4]	63.1	51.5	74.3	56.3
DANN[57]	63.2	51.8	76.8	57.6
JAN[47]	63.9	52.4	76.8	58.3
CDAN[72]	68.4	55.4	80.5	63.8
CDAN+E[72]	70.9	56.7	81.6	65.8
DSAN[48]	73.8	60.6	83.1	67.5
Ours	72.9	60.0	83.6	**67.7**

Table 4.2: Accuracy Comparison of Unsupservised Domain Adaptation on Office-Home Dataset

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
ResNet[25]	74.8 ± 0.3	83.9 ± 0.1	91.5 ± 0.3	78.0 ± 0.2	65.5 ± 0.3	91.2 ± 0.3	80.7
DDC[3]	74.6 ± 0.3	85.7 ± 0.8	91.1 ± 0.3	82.3 ± 0.7	68.3 ± 0.4	88.8 ± 0.2	81.8
DAN[4]	75.0 ± 0.4	86.2 ± 0.2	93.3 ± 0.2	84.1 ± 0.4	69.8 ± 0.4	91.3 ± 0.4	83.3
DANN[57]	75.0 ± 0.4	86.0 ± 0.3	96.2 ± 0.4	87.0 ± 0.5	74.3 ± 0.5	91.5 ± 0.6	85.0
D-CORAL[49]	76.9 ± 0.2	88.5 ± 0.3	93.6 ± 0.3	86.8 ± 0.6	74.0 ± 0.3	91.6 ± 0.3	85.2
JAN[47]	76.8 ± 0.4	88.0 ± 0.2	94.7 ± 0.2	89.5 ± 0.3	74.2 ± 0.3	91.7 ± 0.3	85.8
MADA[62]	75.0 ± 0.3	87.9 ± 0.2	96.0 ± 0.3	88.8 ± 0.3	75.2 ± 0.2	92.2 ± 0.3	85.8
CDAN[72]	76.7 ± 0.3	90.6 ± 0.3	97.0 ± 0.4	90.5.8 ± 0.4	74.5 ± 0.3	93.5 ± 0.4	87.1
iCAN[77]	79.5 ± 0.1	89.7 ± 0.1	94.6 ± 0.2	89.9 ± 0.4	78.5 ± 0.1	92.0 ± 0.1	87.4
DSAN[48]	80.2 ± 0.2	93.3 ± 0.4	97.2 ± 0.3	93.8 ± 0.2	80.8 ± 0.4	95.9 ± 0.4	90.1
Ours	79.8 ± 0.2	93.5 ± 0.2	98.1 ± 0.2	94.4 ± 0.2	79.8 ± 0.1	96.3 ± 0.2	**90.4**

Table 4.3: Accuracy Comparison of Unsupervised Domain Adaptation on ImageCLEF Dataset

Office-31

Office-31 [78] is a very popular dataset for benchmarking domain adaptation. This dataset contains more than 4000 images in 31 classes collected from three different domains: Amazon (A), which consists of images downloaded from amazon.com, and Webcam(W), and DSLR(D), which comprises of images taken by web camera and digital SLR camera with various photographic settings, respectively. Figure 4.3 and 4.4 represent some of the samples from Amazon and Webcam respectively. Table 4.1 reports the performance of our method compared with other works on Office-31 dataset.

Office-Home

Office-Home [79] is another challenging dataset for unsupervised domain adaptation. This dataset contains four domains: Art(Ar), Clipart(CI), Product(Pr), and Real-World(Rw). Each domain has common 65 categories. The Art domain contains the artistic description of objects including painting, sketches etc. The Clipart are the collection of clipart images. In the Product, domain images have no background. The Real-World domain contains an object taken from a regular camera. In Table 4.2, we compare our result with previous methods on Office-Home dataset.

ImageCLEF

ImageCLEF-DA[1] contains three domains: Caltech-256(C), ILSVRC 2012 (I), and Pascal-VOC 2012 (P). Each domain has 12 common classes, and each class has 50 samples. In total, there are 600 images in each domain. Table 4.3 reports the performance of our method with previous methods on ImageCLEF dataset.

[1] http://imageclef.org/2014/adaptation

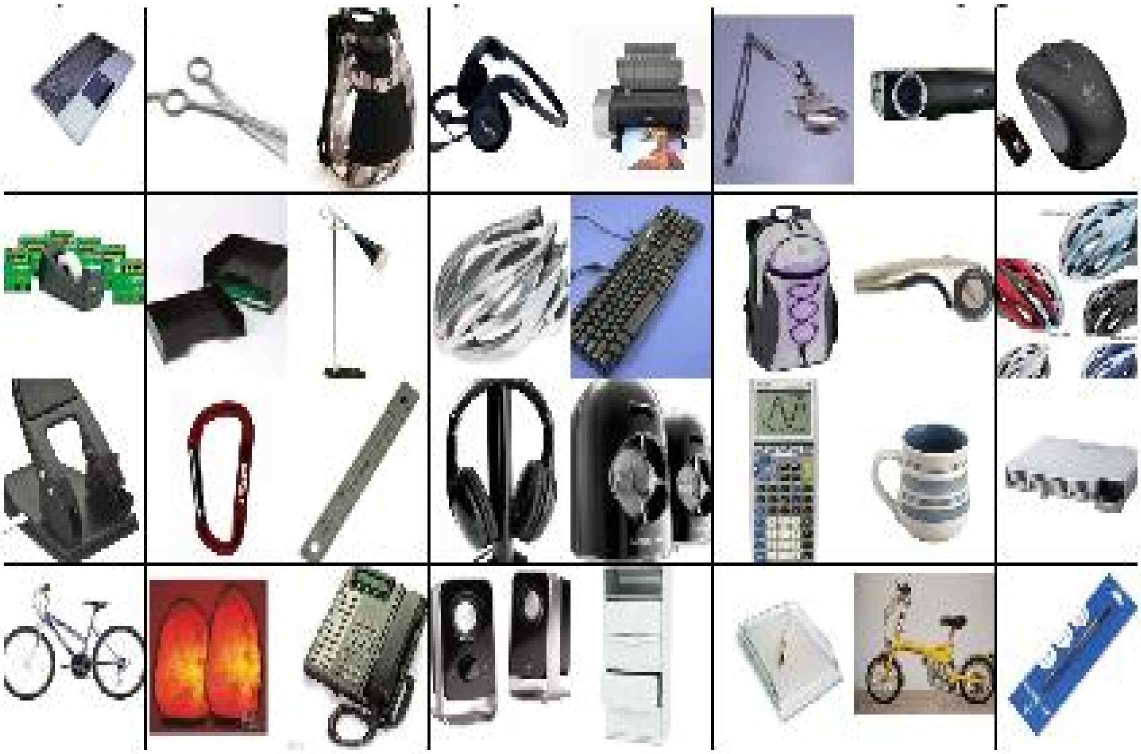

Figure 4.3: Office-31 Dataset on Amazon Pictures

4.4.2 Setup

In our experiment, we used the open-source implementation of a popular deep learning framework, Pytorch [80], to train the models on multiple Nvidia Geforce GTX 1080Ti GPUs. The machine has Intel Core-i7-5930k CPU@ 3.50GHz x 12 processors with 64GB of memory. For the visual classification task, we applied ResNet50 [25] as the backbone network. For comparison, all the baseline models use identical architecture. We fine-tune all the layers except classifier layers from ImageNet[81] pre-trained models and train the fully connected layers for classification through backward-propagation. We set the learning rate to 0.01, batch size to 32, we use stochastic gradient descent (SGD) with a momentum of 0.9, the learning rate is getting changed during SGD using the formula: $lr_{new} = lr_{old}/(1+\alpha(epoch-1)/epoch)^{\beta}$, where $\alpha = 10$, and $\beta = 0.75$.

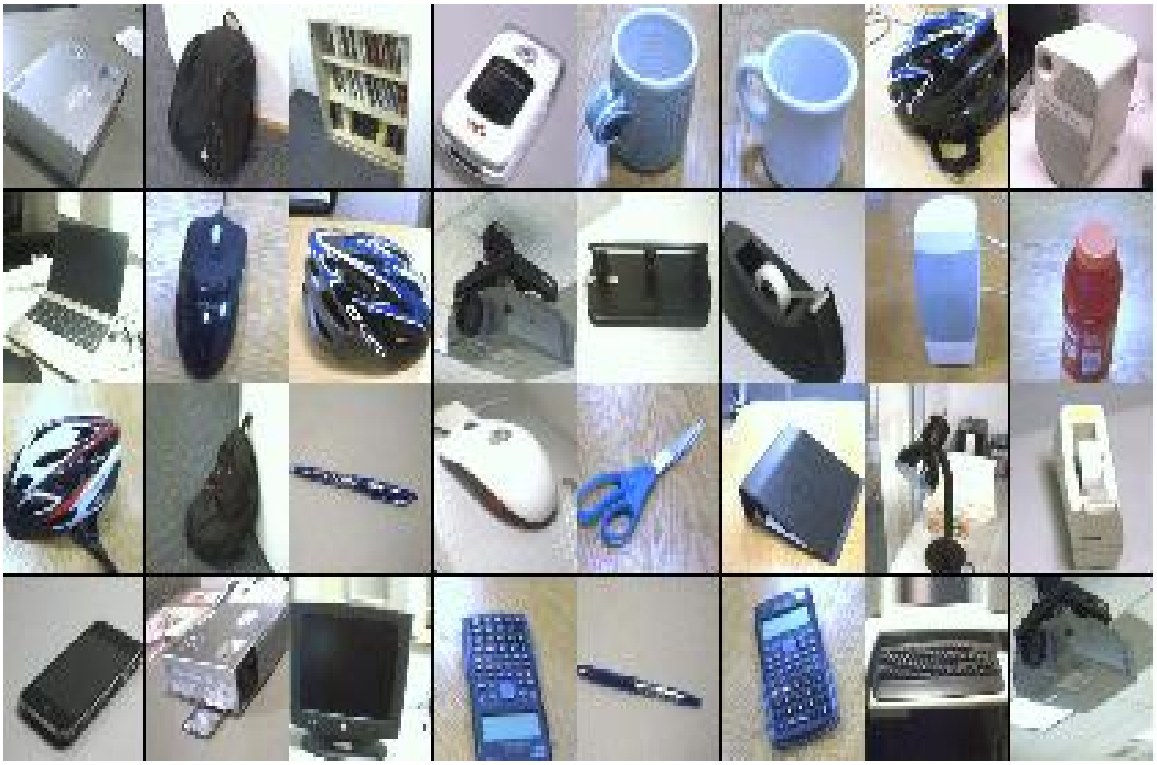

Figure 4.4: Office-31 Dataset on Webcam Pictures

For the adaptation feature norm loss, we observe the embedding size of task-specific features played a major role in norm computation. We found $r = 1$ and $\lambda = 0.05$ provide the best result. the highest value of r is to $R = 5$, so it progresses each step r incrementally. The average classification accuracy and error are reported over three random repeats.

4.5 Results and Discussion

We use our proposed approach for unsupervised subdomain adaptation. We use the protocol to utilize source data with labels and target data without labels. The visual classification results of Office-31, Office-Home, and ImageCLEF-DA are promising. Our method outperforms previous methods on these datasets. Some of the observations from our experiments are:

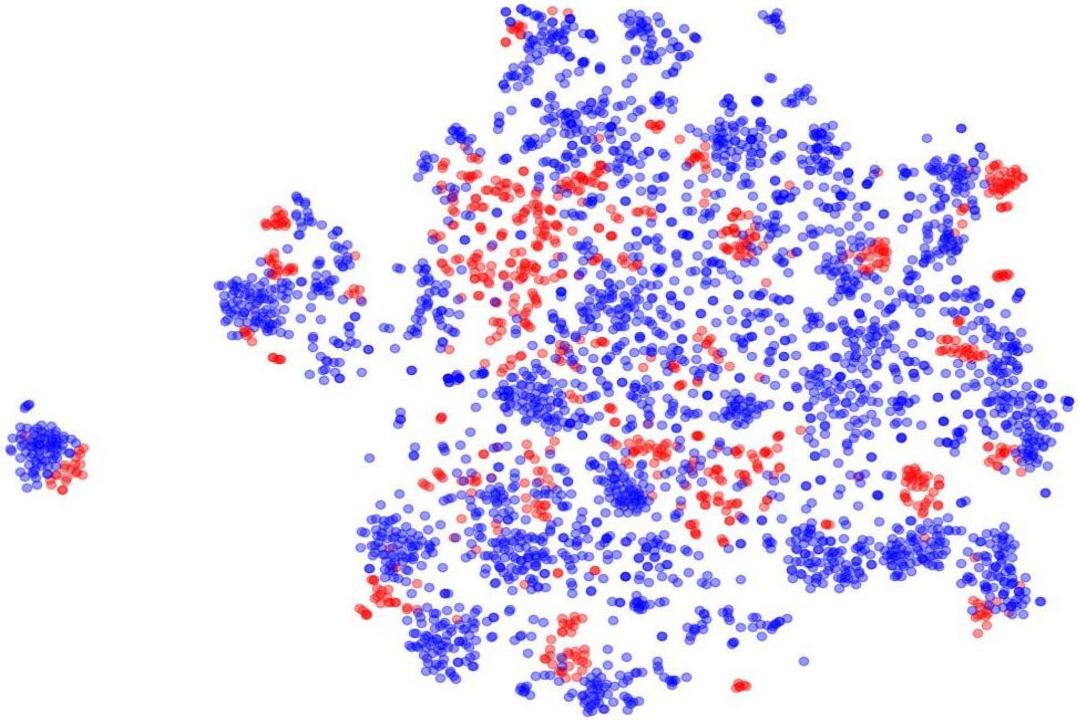

Figure 4.5: t-SNE visualization for Office-31 dataset on task $W \rightarrow A$ (Before Adaptation)

- Comparing our proposed approach with the global domain adaptation methods and several adversarial subdomain adaptation methods [72, 62, 70], these methods are more complex compared to our approach. The reason is most of the methods use the adversarial loss function, and don't consider the kernel mean embeddings between source and target subdomains, and has more number of parameters. Moreover, our method achieves better accuracy compared with others.
- We compared our results with previous adversarial methods [4, 57, 62, 72]. Our results are promising against most of the adversarial techniques. We improved the average accuracy by 2.7% on ImageCLEF dataset, 0.8% on Office-31 dataset,

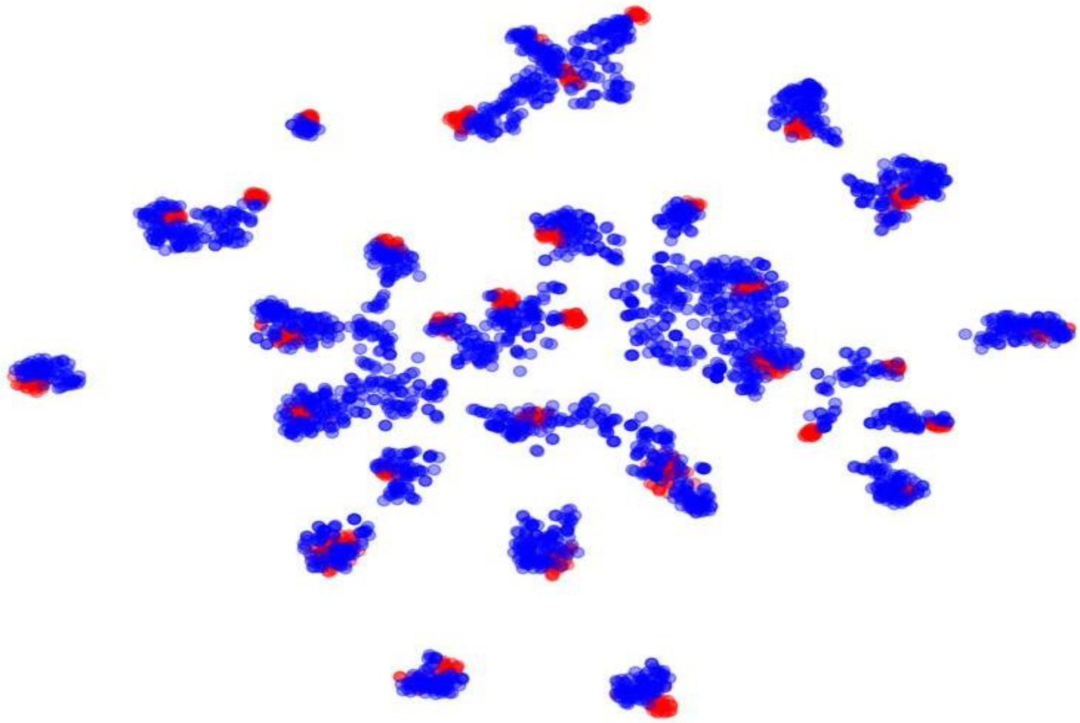

Figure 4.6: t-SNE visualization for Office-31 dataset on task $W \rightarrow A$ (After Adaptation)

and around 2% on Office-Home dataset compared to the most recent adversarial methods [72].

- The t-SNE visualization on the transfer task from Webcam (red) to Amazon (blue) before and after adaptation is presented in Figure 4.5 and Figure 4.6, respectively. In Figure 4.7 and 4.8, t-SNE embeddings for Amazon (red) to Webcam (blue) is showed. Figure 4.9 and 4.10, Figure 4.11 and 4.12, Figure 4.13 and 4.14, and Figure 4.15 and 4.16 shows the before and after domain adaptation of dslr to webcam, webcam to dslr, amazon to dslr, and dslr to amazon respectively.
- We conducted case studies to investigate the sensitivity of parameter Δr in Figure 4.17. It initiallly increases upto $\Delta r = 1$ than gradually decreases.

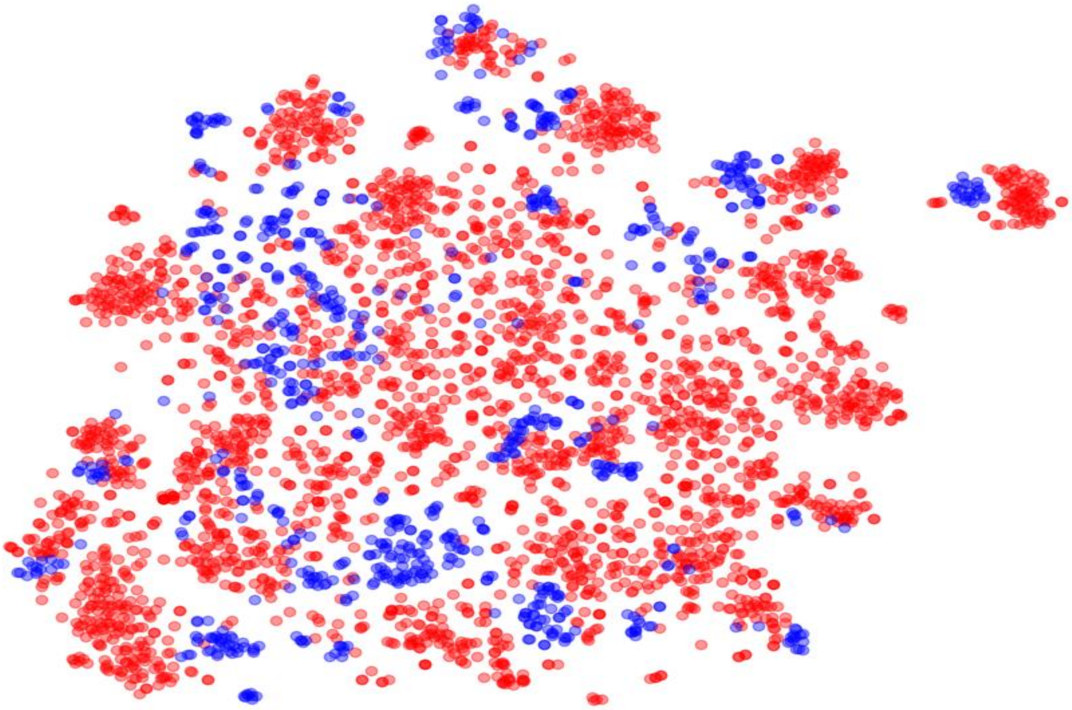

Figure 4.7: t-SNE visualization for Office-31 dataset on task $A \rightarrow W$ (Before Adaptation)

Most of these methods do not consider the subdomain relationship, which effectively captures nuances for each class. Additionally, we incorporate adaptive feature norm loss inside of subdomain distributions. It contributes to apprehend more fine-grained information. The results validate the efficacy of our approach.

4.6 Conclusion

In this work, we have proposed an innovative UDA approach, which incorporates local mean distributed discrepancy measure(LMMD) with adaptive feature norm on subdomain adaptation. Subdomain adaptation can precisely align the distributions of related subdomains within the source and target domains' relevant category. Moreover, subdomain adaptation can boost the transfer gains more if the adaptive feature

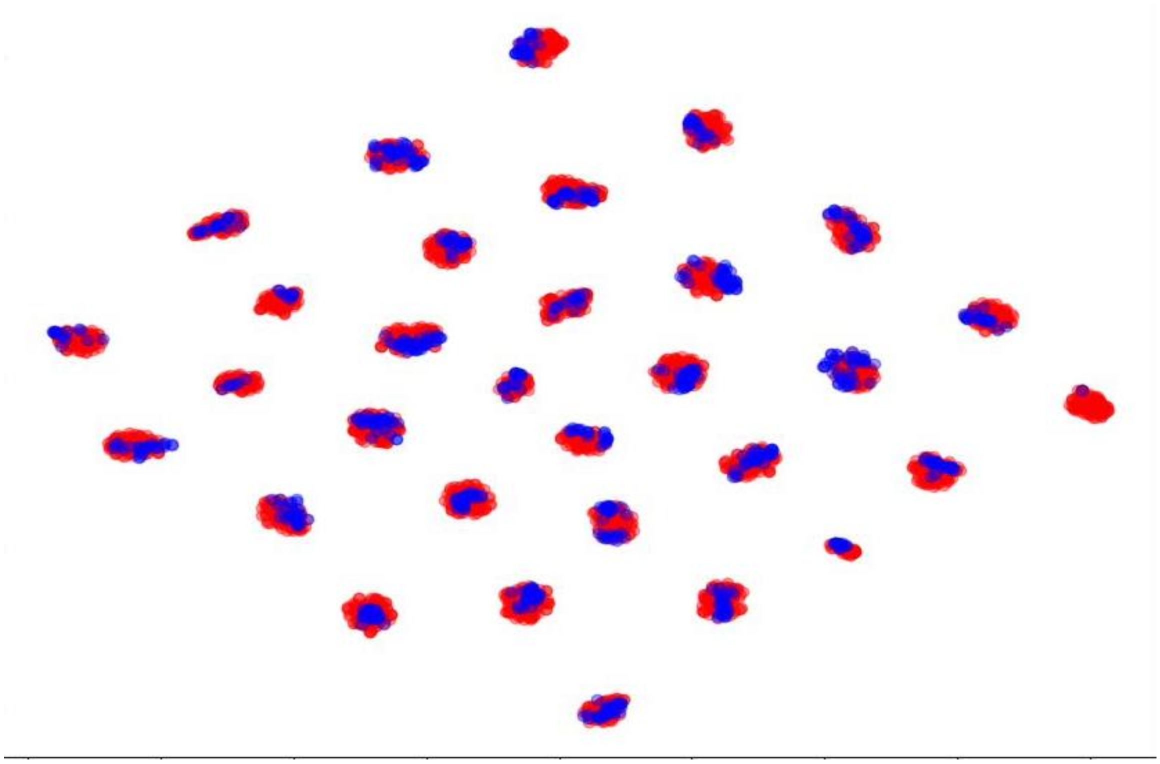

Figure 4.8: t-SNE visualization for Office-31 dataset on task $A \rightarrow W$ (After Adaptation)

norm of the subdomains to a large range of values is added. Extensive experiments were performed on three of the most popular datasets for domain adaptation. Our results show the method's effectiveness, implying that task-specific features with larger norms are more transferable on subdomain adaptation. In future, we will investigate adaptive feature norm on an adversarial based unsupervised domain adaptation scenario.

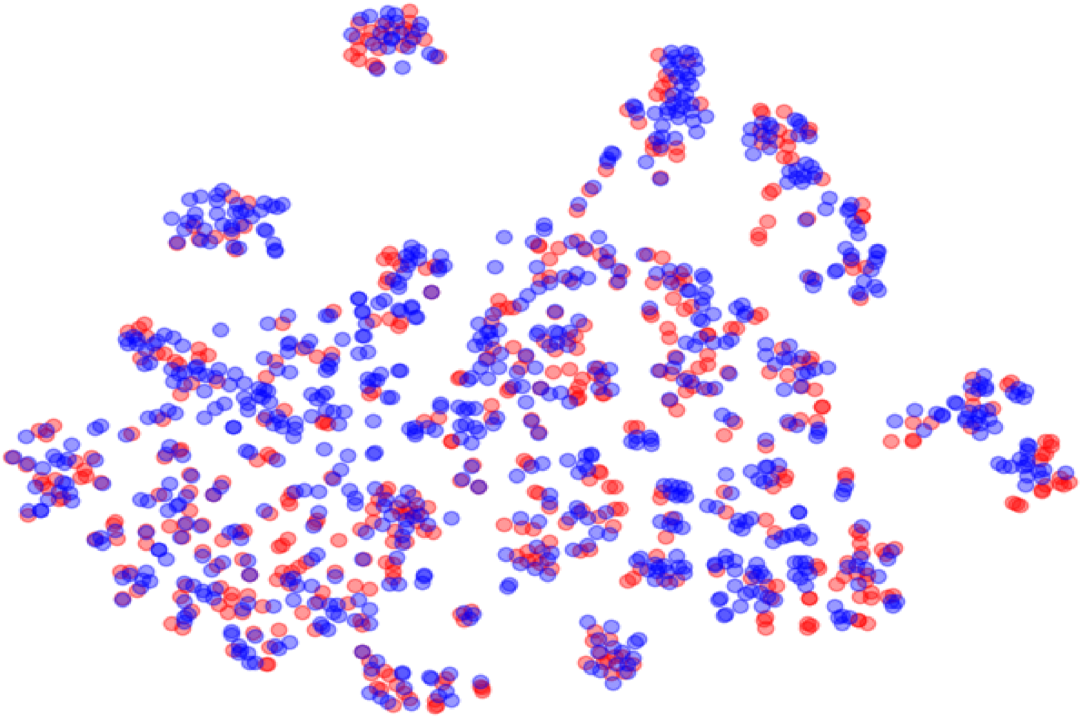

Figure 4.9: t-SNE visualization for Office-31 dataset on task $D \rightarrow W$ (Before Adaptation)

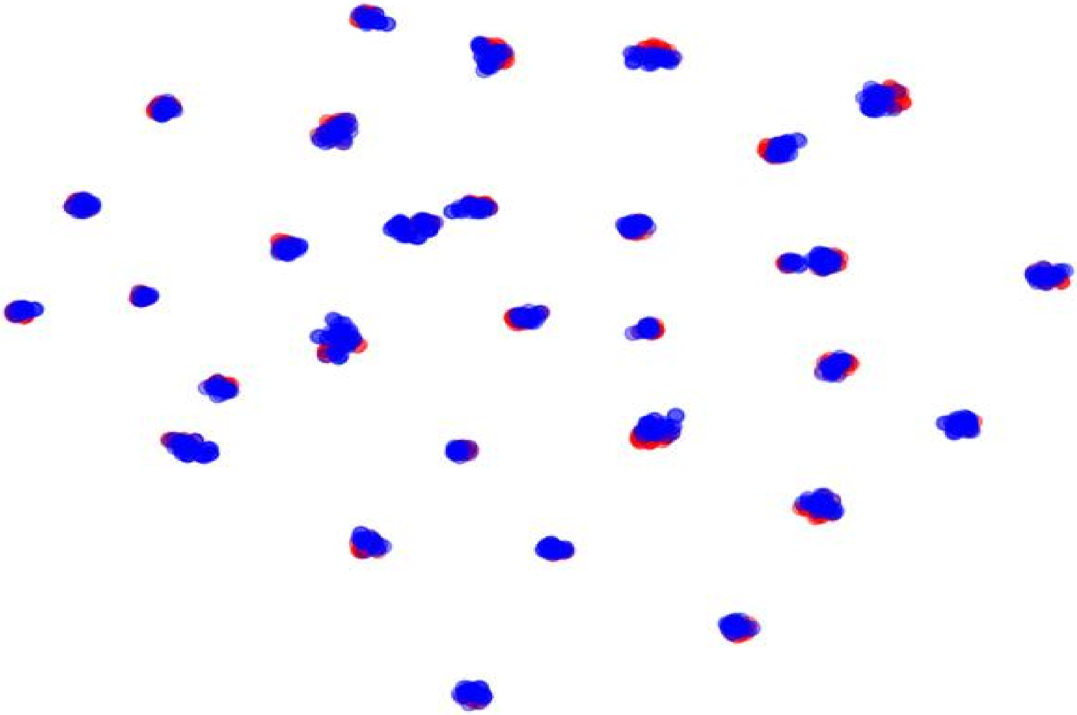

Figure 4.10: t-SNE visualization for Office-31 dataset on task $D \rightarrow W$ (After Adaptation)

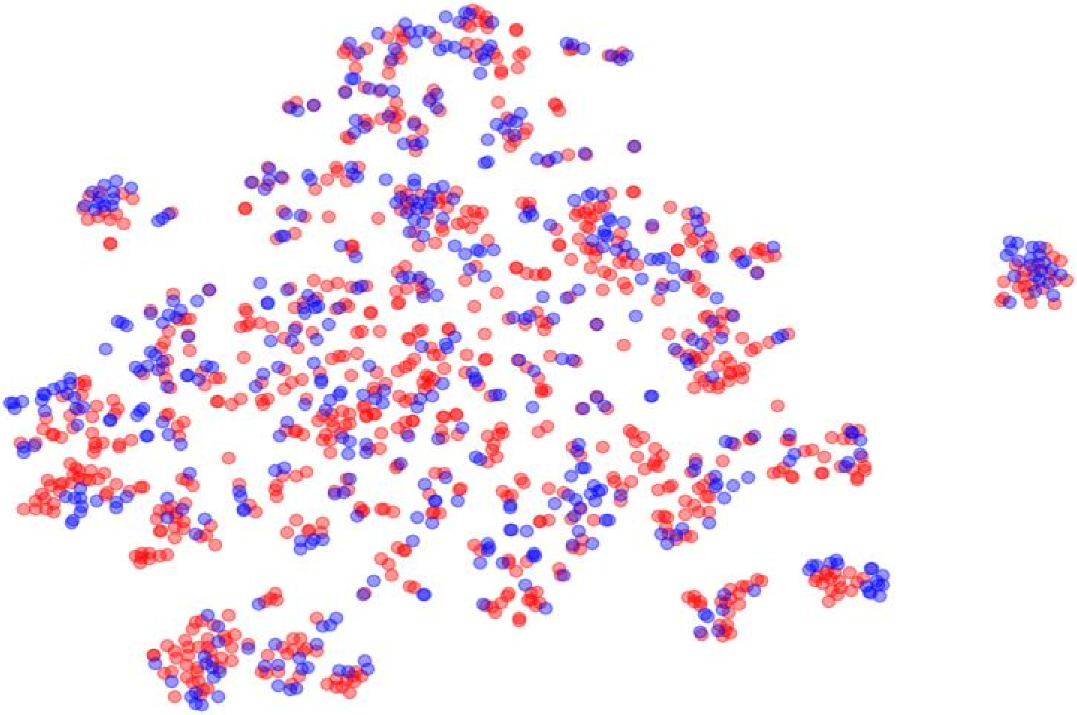

Figure 4.11: t-SNE visualization for Office-31 dataset on task $W \rightarrow D$ (Before Adaptation)

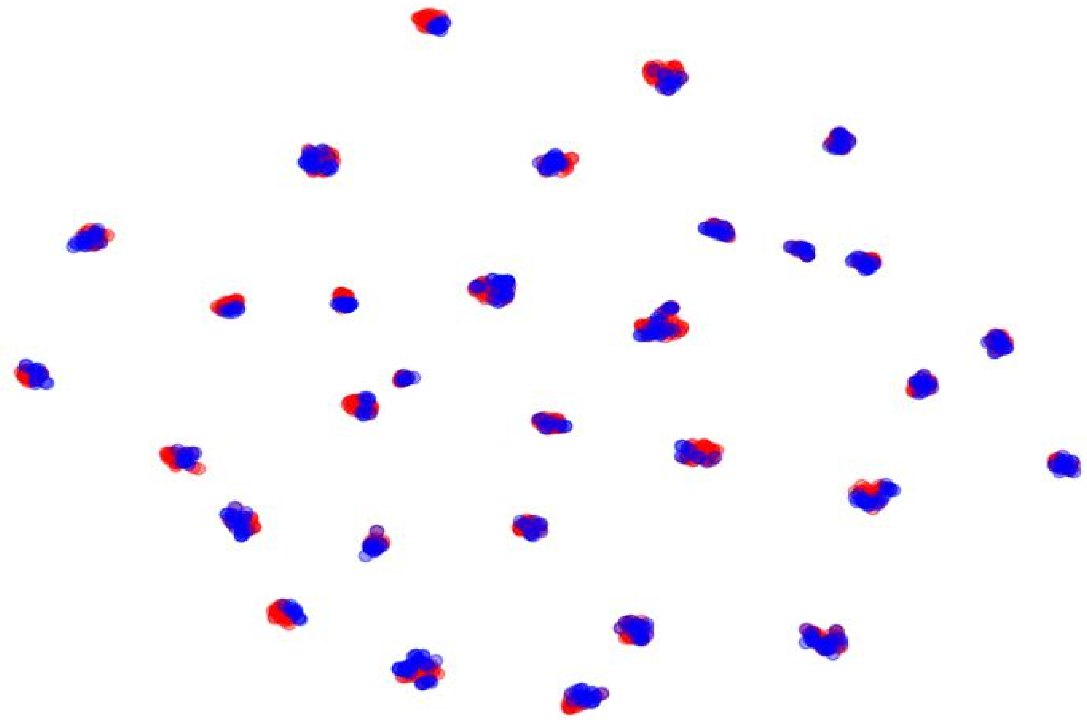

Figure 4.12: t-SNE visualization for Office-31 dataset on task $W \rightarrow D$ (After Adaptation)

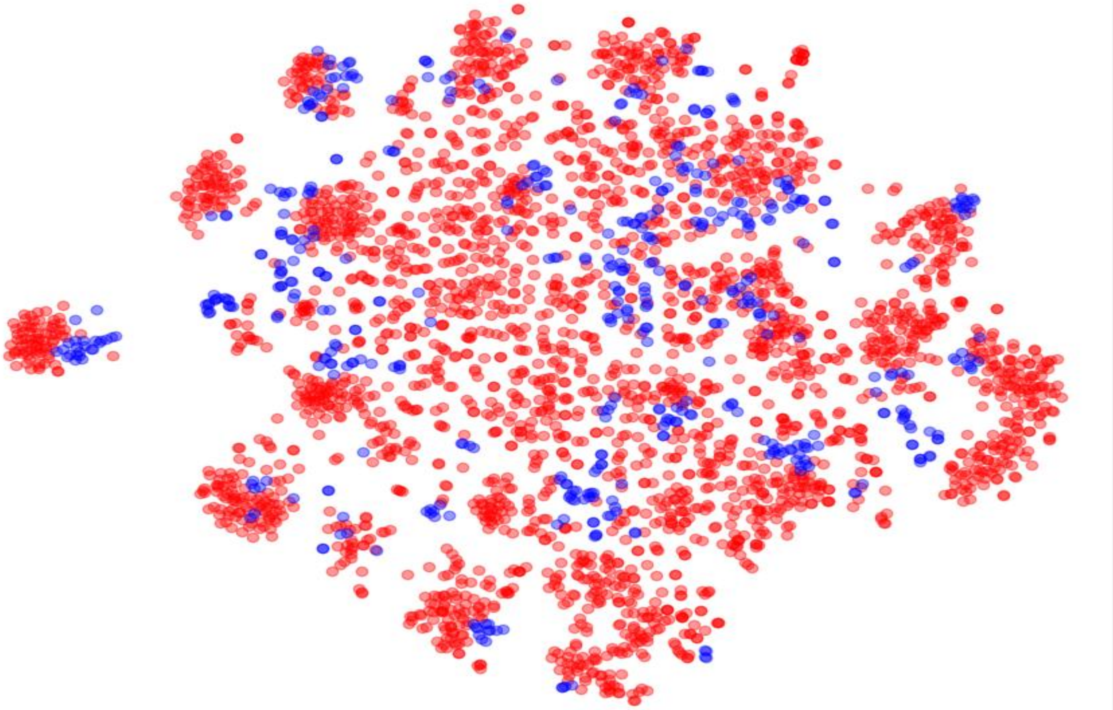

Figure 4.13: t-SNE visualization for Office-31 dataset on task $A \rightarrow D$ (Before Adaptation)

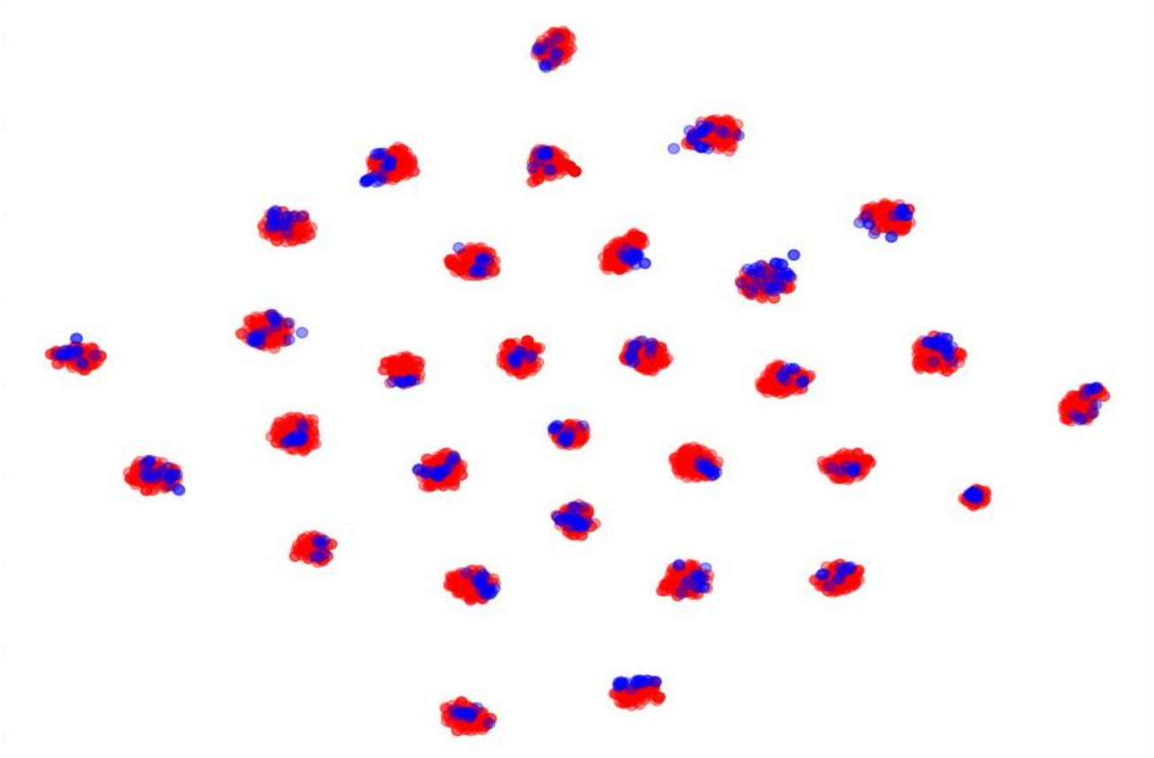

Figure 4.14: t-SNE visualization for Office-31 dataset on task $A \rightarrow D$ (After Adaptation)

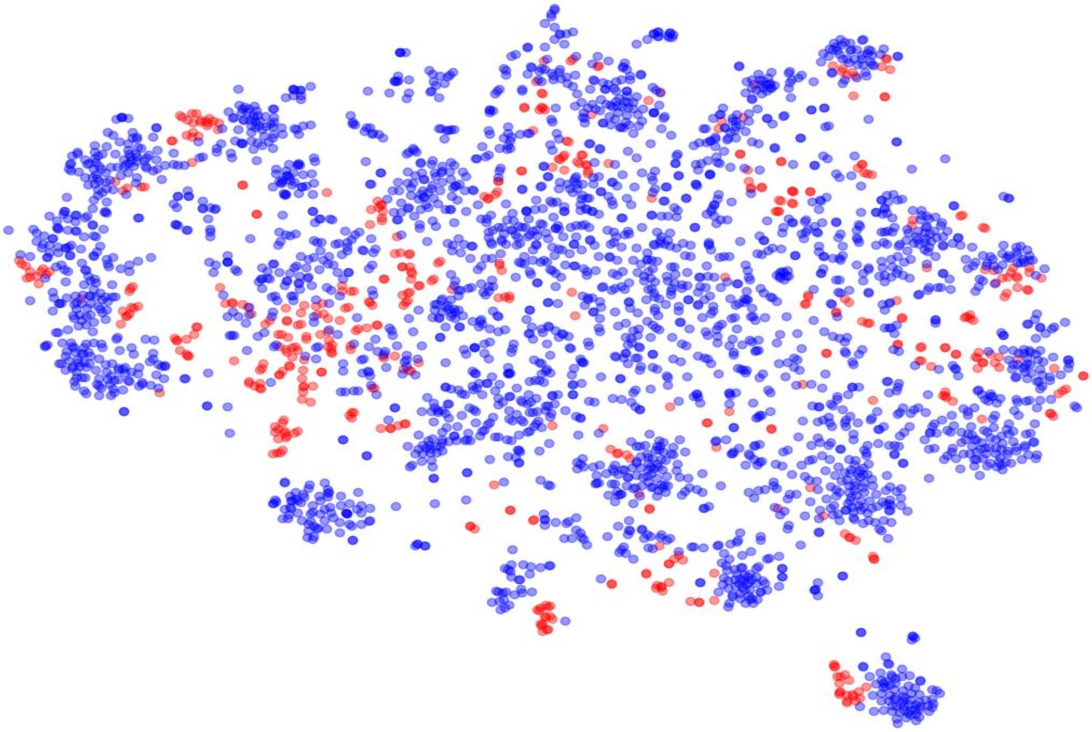

Figure 4.15: t-SNE visualization for Office-31 dataset on task $D \rightarrow A$ (Before Adaptation)

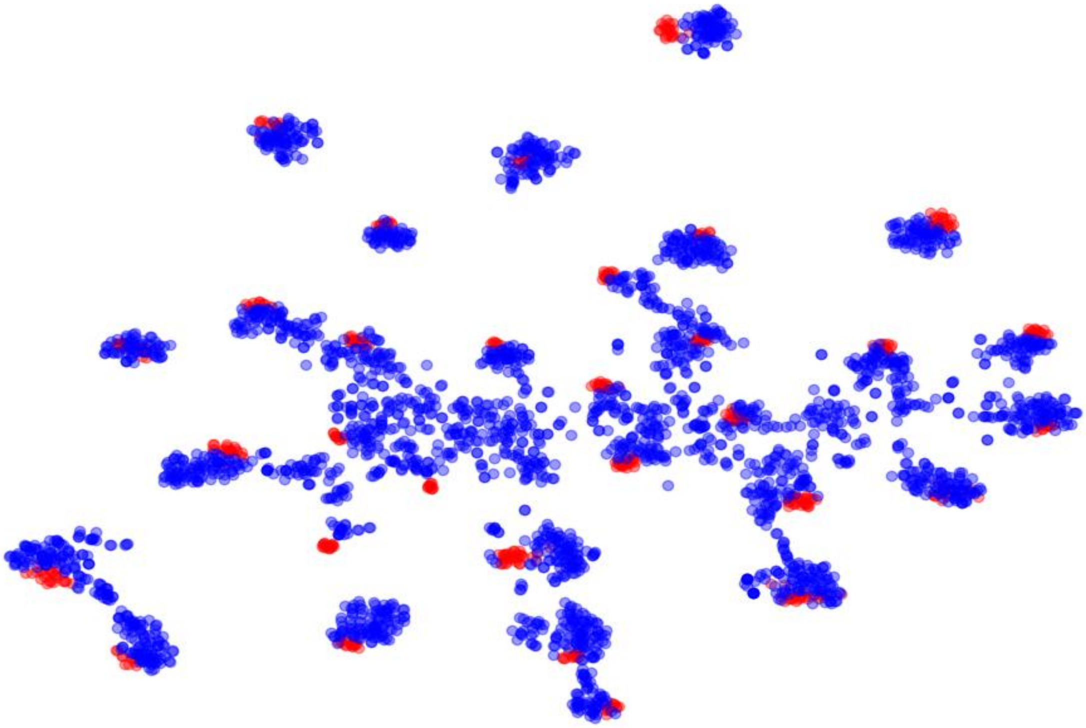

Figure 4.16: t-SNE visualization for Office-31 dataset on task $D \rightarrow A$ (After Adaptation)

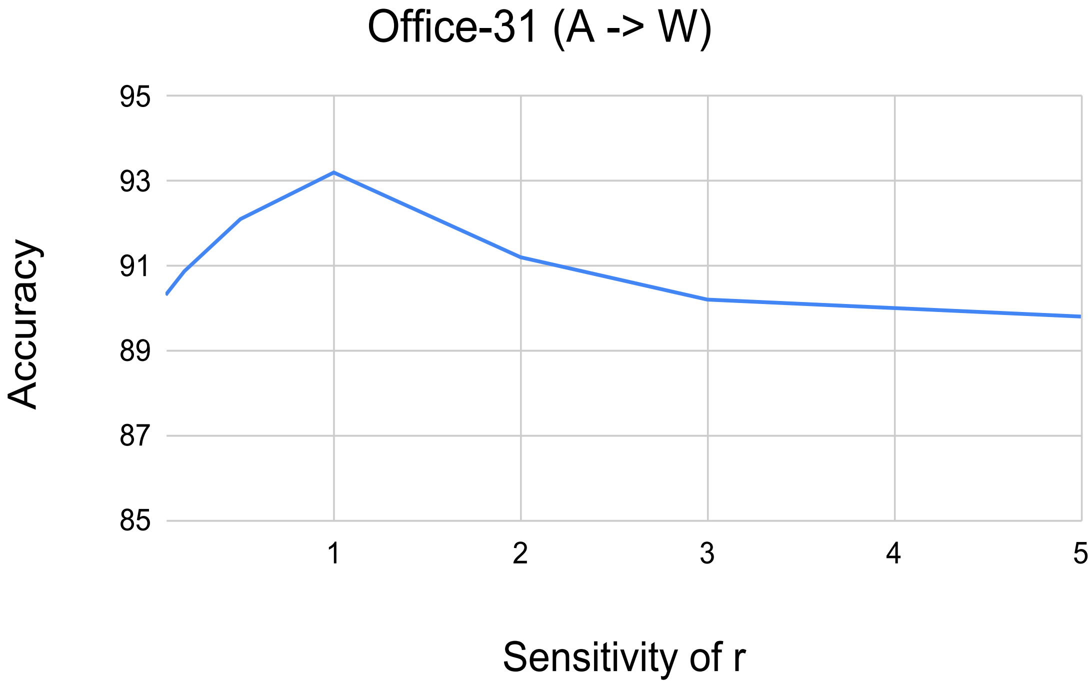

Figure 4.17: parameter sensitivity of Δr on Office-31 (Amazon→Webcam) dataset

CHAPTER 5

DISCUSSION

Transfer Learning is one of the most widely applied methods in the field of computer vision. Typically, a machine learning model is trained first on a large labeled dataset such as ImageNet [81], and then fine-tuned on a target dataset. During fine-tuning, the parameters for the network layers except the classification layer are initialized from the pre-trained model. This method has proven very effective for object recognition [32, 82], and many other applications. This method has been leveraged to tackle applications where dataset size is limited. In this dissertation, we demonstrated the effect of transfer learning on fine-grained visual categorization. We show how the choice of pre-training data can play a crucial role when fine-tuned on a new dataset. We have shown that more pre-training data does not always help. Additionally, matching target domain distribution can improve the accuracy of the model.

One of the significant problems in real-world applications is the distribution change and domain shift between the source and target domain. Domain adaptation is a particular case of transfer learning that exploits labeled data in one or more relevant source domains to execute new tasks in the target domain. Deep neural network-based domain adaptation approaches have shown remarkable results in visual categorization applications. Unsupervised domain adaptation, where no labeled data in the target domain is available in training, has enticed the research community. In this dissertation, we investigate the unsupervised domain adaptation method on visual application tasks among different domains. Our method incorporates an unsupervised sub-domain adaptation technique with adaptive feature norm alignment to achieve state-of-the-art results on several datasets.

Domain adaptation using the deep neural network has been utilized successfully in visual classification applications such as object recognition, object detection, face recognition, semantic segmentation, and person re-identification. However, most methods have yet to achieve decent results on these tasks with no data or minimal data. Various techniques [56] such as domain-invariant feature learning, domain mapping, statistics-based methods have been developed for achieving good performance. Some problems still can be explored to further improvements.

- **Bi-Directional Adaptation**: Domain adaptation addresses some problems, which shows decent results in one way, but not in two ways. For example, SVHN → MNIST shows accuracy above 90%, while for the reverse case of MNIST → SVHN, the highest accuracy is around 80%. This accuracy shows how the bi-directional adaptation method is much harder [10, 56]. More work is needed to strengthen the bi-directional adaptation.
- **Combining Methods**: Domain adaptation methods typically align feature distributions; another way of the research aligns the joint or conditional distribution [72, 83, 84] of the feature and label spaces instead. Some methods found aligning jointly or conditionally improves adaptation performance when handling multi-modal data distributions [72]. Other domain adaptation strategies can apply joint or conditionally distribution to boost domain adaptation performance instead of depending on feature distribution.
- **Multiple Source or Multiple Target based Domain Adaptation**: Most of the existing methods focus on single-source homogeneous domain adaptation. Multiple source domains or multiple target domains can provide additional gains in performance by employing more data. In the area of heterogeneous feature

spaces or various other levels of supervision such as semi-supervised [85] domain adaptation may boost the performance gains. This area can be explored more in the future.

Some of the methods show promising results, but further research can be done on novel method combinations, direct comparisons, domain generalization, and new datasets and applications.

CHAPTER 6

CONCLUSION

This dissertation investigates the domain adaptive transfer learning approach on fine-grained visual categorization (FGVC), provided a comprehensive literature study on unsupervised domain adaptation on visual categorization, and an innovative stepwise adaptive feature norm-based approach for unsupervised subdomain adaptation. The contributions in this work are as follow:

1. We demonstrated a domain adaptive transfer learning approach that combines with visual attention-based data augmentation and can achieve state-of-the-art results on FGVC datasets.
2. We presented the relationship of top-1 accuracy and domain score on six commonly used FGVC datasets. Additionally, we illustrated the effect of image resolution in transfer learning in detail.
3. We presented a survey of different types of existing techniques on unsupervised domain adaptation for visual categorization.
4. An innovative stepwise adaptive feature norm-based approach for unsupervised subdomain adaptation was proposed. This approach employs progressively learning task-specific features, aligning relevant subdomains in unsupervised scenarios.
5. We compared results with both adversarial and non-adversarial methods to show the efficacy of the work.

REFERENCES

[1] T. Hu and H. Qi, "See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification," *arXiv preprint arXiv:1901.09891*, 2019.

[2] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[3] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[4] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning.* PMLR, 2015, pp. 97–105.

[5] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[6] L. Zhang and X. Gao, "Transfer adaptation learning: A decade survey," *arXiv preprint arXiv:1903.04687*, 2019.

[7] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4109–4118.

[8] Z. Ge, A. Bewley, C. McCool, P. Corke, B. Upcroft, and C. Sanderson, "Fine-grained classification via mixture of deep convolutional neural networks," in *2016*

IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016, pp. 1–6.

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[10] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning.* PMLR, 2015, pp. 1180–1189.

[11] A. Imran and V. Athitsos, "Domain adaptive transfer learning on visual attention aware data augmentation for fine-grained visual categorization," in *International Symposium on Visual Computing.* Springer, 2020, pp. 53–65.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition.* Ieee, 2009, pp. 248–255.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[15] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing.* IEEE, 2008, pp. 722–729.

[16] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[17] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.

[18] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fgvc: Stanford dogs," in *San Diego: CVPR Workshop on FGVC*, vol. 1, no. 2, 2011.

[19] P. Zhang, Y. Zhong, Y. Deng, X. Tang, and X. Li, "A survey on deep learning of small sample in biomedical image analysis," *arXiv preprint arXiv:1908.00473*, 2019.

[20] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[21] X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas, "Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2226–2234.

[22] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1–10.

[23] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Cvae-gan: fine-grained image generation through asymmetric training," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2745–2754.

[24] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1143–1151.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[27] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear cnn models for fine-grained visual recognition,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[28] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, “Adversarial complementary learning for weakly supervised object localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1325–1334.

[29] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 842–850.

[30] J. Fu, H. Zheng, and T. Mei, “Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4438–4446.

[31] H. Zheng, J. Fu, T. Mei, and J. Luo, “Learning multi-attention convolutional neural network for fine-grained image recognition,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5209–5217.

[32] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: an astounding baseline for recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[33] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning*, 2014, pp. 647–655.

[34] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[35] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[36] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of transferability for a generic convnet representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1790–1802, 2015.

[37] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.

[38] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[39] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8769–8778.

[40] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[41] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2403–2412.

[42] T.-Y. Lin and S. Maji, "Improved bilinear pooling with cnns," *arXiv preprint arXiv:1707.06772*, 2017.

[43] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a cnn for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4148–4157.

[44] P. Li, J. Xie, Q. Wang, and Z. Gao, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 947–955.

[45] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 801–814, 2018.

[46] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2200–2207.

[47] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *International conference on machine learning.* PMLR, 2017, pp. 2208–2217.

[48] Y. Zhu, F. Zhuang, J. Wang, G. Ke, J. Chen, J. Bian, H. Xiong, and Q. He, "Deep subdomain adaptation network for image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[49] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision.* Springer, 2016, pp. 443–450.

[50] F. Zhuang, X. Cheng, P. Luo, S. J. Pan, and Q. He, "Supervised representation learning: Transfer learning with deep autoencoders," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[51] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3723–3732.

[52] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," *arXiv preprint arXiv:1608.06019*, 2016.

[53] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European Conference on Computer Vision.* Springer, 2016, pp. 597–613.

[54] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International Conference on Machine Learning.* PMLR, 2017, pp. 1857–1865.

[55] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[56] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–46, 2020.

[57] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural net-

works," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[58] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4893–4902.

[59] P. O. Pinheiro, "Unsupervised domain adaptation with similarity learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8004–8013.

[60] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2107–2116.

[61] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," *Advances in neural information processing systems*, vol. 29, pp. 469–477, 2016.

[62] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proceedings of the IEEE conference on AAAI*, 2018, pp. 3934–3941.

[63] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6502–6509.

[64] J. Q. Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, "Dataset shift in machine learning," *The MIT Press*, vol. 1, p. 5, 2009.

[65] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, 2011, pp. 1521–1528.

[66] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[67] Y. Zhu, F. Zhuang, and D. Wang, "Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5989–5996.

[68] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[69] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.

[70] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *International Conference on Machine Learning*, 2018, pp. 5423–5432.

[71] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8503–8512.

[72] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, 2018, pp. 1640–1650.

[73] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1426–1435.

[74] J. Ye, X. Lu, Z. Lin, and J. Z. Wang, "Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers," *arXiv preprint arXiv:1802.00124*, 2018.

[75] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[76] A. Kumar, P. Sattigeri, K. Wadhawan, L. Karlinsky, R. Feris, W. T. Freeman, and G. Wornell, "Co-regularized alignment for unsupervised domain adaptation," *arXiv preprint arXiv:1811.05443*, 2018.

[77] W. Zhang, W. Ouyang, W. Li, and D. Xu, "Collaborative and adversarial network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3801–3809.

[78] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European conference on computer vision.* Springer, 2010, pp. 213–226.

[79] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5018–5027.

[80] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[81] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[82] J. Ngiam, D. Peng, V. Vasudevan, S. Kornblith, Q. V. Le, and R. Pang, "Domain adaptive transfer learning with specialist models," *arXiv preprint arXiv:1811.07056*, 2018.

[83] S. Cicek and S. Soatto, "Unsupervised domain adaptation via regularized conditional alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1416–1425.

[84] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty, "Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 447–463.

[85] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8050–8058.

BIOGRAPHICAL STATEMENT

Ashiq Imran was born in Bangladesh. He received his Bachelor's degree in Computer Science and Engineering from the Bangladesh University of Engineering and Technology in 2013. He received his Master's degree in Computer Science from North Carolina A&T State University in 2015. In Fall 2016, he started his PhD journey at the University of Texas at Arlington in the department of computer science. His current research interests include Deep Learning, Applied Machine Learning, Computer Vision and Data Science. He conducted research on developing domain adaptive transfer learning techniques that reduce the dependency for large-scale annotated data to train machine learning models from scratch as well as adapting to new domains by transfer learning from source domain to target domain.